INTEGRATION OF GIS AND MACHINE LEARNING TECHNIQUES TO INVESTIGATE THE IMPACT OF ENVIRONMENTAL CONTEXTS ON TRAVEL MODES

BY

KANGJAE LEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Mei-Po Kwan, Chair
Professor Shaowen Wang
Associate Professor Feng Liang
Assistant Professor Matthew Browning

ABSTRACT

Related to the promotion of physical activity, a growing body of research has adopted the definition of active travel modes. Active travel modes have made a great contribution to overall physical activity and, therefore, it is important to understand the active travels associated with environmental facilitators or barriers in physical activity and transportation research. Residential neighborhoods around individuals' home locations were a primary focus in previous studies to examine the associations between active travels and environmental factors and, for the last decade, researchers have begun using global positioning system (GPS) trajectories of individuals to consider their daily paths for actual exposure estimation to various environments. Empirical findings in the existing studies, however, showed inconsistent outcomes of the associations. In addition, more advanced analytical approaches have not yet been explored, regardless of a large amount of GPS trajectories in hand, which have great potential to find more valuable and various outcomes. Thus, this study seeks to provide comprehensive data-driven approaches to further investigate the associations between travel modes and environmental contexts using the geographic information system (GIS) and machine learning techniques. An automatic travel mode classification algorithm is developed using GPS and accelerometer data to advance travel mode detection in health and transportation research. When it comes to exposure estimation to various environments, this study focuses on buffer analysis, which has been widely used in previous studies, and examines how distance, as one of the buffer characteristics, can affect findings of the associations between travel modes and environmental factors to give insights into accurate estimation of immediate surroundings along the daily trajectories of individuals. In addition, a novel framework is proposed and adopted to perform mapping of travel modes and explore complex contextual influences on travel modes at different levels of scales using

machine learning models. In the era of big data, this dissertation suggests methodological directions for various fields of study to adequately deal with a large quantity of sensor data collected from many participants, derive informative measures for classifying health behaviors from the sensor data, and conduct exploratory analyses and produce meaningful knowledge using machine learning models with GIS data.

ACKNOWLEDGEMENTS

Firstly, I would like to express my deep appreciation to my advisor Prof. Mei-Po Kwan, for guiding me with her expertise and kindly supporting me throughout my Ph.D. studies for the last five years at the University of Illinois at Urbana-Champaign (UIUC). I am honored and feel very fortunate to have a great learning opportunity to work under her guidance.

I am thankful to my thesis committee: Prof. Shaowen Wang, Prof. Feng Liang, and Prof. Matthew Browning for providing me intuitive and valuable comments from various perspectives to further enrich my dissertation research. Particularly, I extend my sincere gratitude to Prof. Matthew Browning for his above and beyond support as I continue my Ph.D. training at UIUC.

My earnest appreciation also goes to Prof. Barbara Minsker, Prof. Ming Kuo, Prof. Daniel Miller, Prof. Alessandro Rigolon, and Dr. Pushpendra Rana who provided me opportunities to join their research projects as a research assistant. With their precious support, I was able to gain invaluable experience in various research fields.

I am grateful to Prof. Linda Smith, Prof. Jodi Schneider, Prof. David Dubin, Prof. Kate McDowell, Prof. Matthew Turk, and Prof. Nicolas LaLone in iSchool at my current university for offering me wonderful teaching opportunities to work as a teaching assistant and learn from their great pedagogical methods.

Lastly, I cannot thank enough Dr. Renear and Karin in Illinois Informatics, whose support and guidance have been always with me since I started pursuing my Ph.D. degree in Illinois Informatics. For their help, I could continue my study and nurture my enthusiasm for interdisciplinary research.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

PA: Physical Activity

GPS: Global Positioning System

GIS: Geographic Information System

UGCoP: Uncertain Geographic Context Problem

ATM: Active Travel Mode

SVM: Support Vector Machine

RF: Random Forest

XGB: eXtreme Gradient Boosting

HDOP: Horizontal Dilution Of Precision

CRHTI: Chicago Regional Household Travel Inventory

XSEDE: Extreme Science and Engineering Discovery Environment

SMAIN: Spatio-temporal Mapping And INterpretation

LIME: Local Interpretable Model-agnostic Explanations

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 INTRODUCTION

Physical activity (PA) has drawn the attention of researchers as an influential factor on human health. PA is categorized as light, moderate, and vigorous in terms of its intensity. In particular, performing regular moderate to vigorous PA, such as brisk walking and jogging, helps to decrease risks of physical and mental health problems, including cardiovascular diseases, obesity, type II diabetes, anxiety, and depression (Fox, 1999; Gordon-Larsen et al., 2006; Physical Activity Guidelines Advisory Committee, 2008; Wei et al., 2000). According to the World Health Organization (2011), adults aged 18–64 are recommended to engage in moderate PA for at least 150 minutes or vigorous PA for 75 minutes per week to gain health benefits.

In PA research, as tracking technology advances, global positioning system (GPS) devices have been widely adopted to record daily movements of individuals, combined with moderate to vigorous PA, to objectively detect places people visit when they engage in the high intensity of PA (Troped et al., 2010; Almanza et al., 2012; Jansen et al., 2016). For example, Rodríguez et al. (2012) used GPS observations with buffer analysis, as a geographic information system (GIS) analysis method, to estimate the daily exposure of adolescent females to built environment characteristics associated with their moderate to vigorous PA. The objective detection of visited places using GPS data aids in accurate estimation of the actual influence of environments on people's PA in not only residential neighborhoods but also other important areas farther than the home locations, which may greatly affect people's PA. The importance of surrounding environments in non-residential areas, along with residential neighborhoods, to people's health behaviors and outcomes was highlighted by recent studies (Diez, Roux, & Mair,

2010; Perchoux et al., 2013). GPS trajectories representing individuals' continuous trips, collected from GPS receivers with high-resolution spatial and temporal information, particularly help to mitigate the uncertain geographic context problem (UGCoP) (Kwan, 2012a, b, 2013, 2018b). UGCoP, as a methodological issue, refers to the notion that the geographical delineation of contextual units or neighborhoods can influence the effects of area-based attributes (e.g., percentage of green space) on individual behaviors or outcomes (e.g., PA).

Related to the promotion of PA, a growing body of research has adopted the definition of active travel modes (ATMs) (Oliver et al., 2015; Duncan et al., 2016; Helbich et al., 2016; Voss et al., 2016). Compared to inactive travel modes such as private cars and public transit, ATMs are referred to as a subset of PA, including walking and biking. ATMs make a great contribution to overall PA; therefore, it is important to understand the active travels associated with environmental factors for PA and transportation research. According to Gordon-Larsen and colleagues (2005), the high utilization of active travels for commuting is mostly exhibited among the group of young adults who meet PA recommendation — 150 minutes of moderate to vigorous PA per week. Young adults who don't meet the PA recommendation are correlated with the high utilization of inactive travel modes to workplaces or school. Therefore, the contribution of active travels to the increase in people's PA levels is immense, and needs to be thoroughly understood, as associated with influential factors in physical and social contexts.

Empirical findings in previous studies, however, have shown inconsistent outcomes between PA and environmental factors (Hoehner et al., 2005; McGinn et al., 2007; Nagel et al., 2008; Troped et al., 2010; Boruff, Nathan, & Nijënstein, 2012). For instance, Troped et al. (2013) found that frequent use of green spaces and access to recreational facilities, like trails, in neighborhoods showed positive associations with adults' PA, whereas Hoehner et al.'s findings

(2005) indicated that there is no significant association of adults' utilitarian (e.g., walking for errands and commuting to work) or recreational (e.g., biking during leisure time) PA with places to exercise, nearby parks, trails, and private fitness facilities. Such inconsistency does not provide supportive evidence for ecological models that the environment has influence on human behaviors, including PA, and rather, brings us difficulties in understanding expectations of specific environmental contexts for promotion of active living (Sallis et al., 2006). Ecological models as underpinning theories have been introduced to explain the interactions between PA and various factors at multiple levels of environments, including social and physical environment. In this regard, there is a gap between what the foundational framework depicts regarding the interactions and mixed findings of studies in practice, which needs to be bridged by more in-depth investigation and exploration of the associations.

1.2 RESEARCH QUESTIONS

This study seeks to provide a comprehensive data-driven approach to further investigating the associations between travel modes and environmental contexts using GIS and machine learning techniques. It suggests systematic approaches for the use of GPS data in PA and transportation research in various ways. Since GPS trajectories collected from individuals have a large amount of observations with spatial and temporal information, a novel data-driven approach to the estimation of dynamic interactions between individuals and environments, when engaged in certain types of travel modes, especially needs to be explored. When it comes to exposure estimation to various environments, this study focuses on buffer analysis, which has been widely used in previous studies and is an appropriate method to solely consider actual spatial and temporal dynamics represented in recorded GPS trajectories. This study particularly

looks at variations in the impact of environmental factors as different scales (e.g., global vs. local) and time dimensions that might help discover plausible associations between travel modes and certain environmental contexts. Thus, the main hypotheses of this study are 1) different sizes of buffers to estimate spatially immediate and temporally momentary exposures have an impact on the strength and significance of the associations between ATMs and environmental contexts and 2) the impact of environmental settings on ATMs varies in line with different scales and time points. Specific research questions and objectives in accordance with the proposed hypotheses are presented as follows:

- Research question 1: How can ATMs be identified using GPS data?
    - Develop a travel mode classification algorithm using machine learning techniques
    - Propose an approach to adopting the focal variation concept of derived metrics in GPS trajectories for the accurate identification of bicycle mode and the status when people are in vehicles
    - Explore the relative importance of measures derived from GPS trajectories in classifying travel modes
- Research question 2: How does the association between ATMs and environmental factors vary depending on the characteristics of the buffer?
    - Conduct sensitivity analyses to investigate varying associations between travel modes and environmental factors
    - Explore the varying associations and statistical significance associated with the distance of the buffer to find reliable characteristics

    - Calculate measures and investigate the associations between ATMs and environmental factors using statistical analyses with buffer analysis, with a resulted reliable distance
- Research question 3: How can the spatio-temporal patterns of the ATMs be predicted and visually explored across a city area using a massive amount of GPS points?
    - Propose a novel approach to map the predicted results and explore the patterns of the active travels with different environmental context layers
    - Generate predictive models using support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGB) with multiple predictors related to various environmental contexts
    - Compare the performance and predicted results of the three machine learning models and investigate their validation
    - Explore spatio-temporal patterns of environmental factors and interpret their contribution to the predictions of travel modes

1.3 BACKGROUND

1.3.1 Ecological models and empirical evidence

To account for the associations between health behavior and various environmental contexts, ecological models have been adopted as theoretical foundations. Ecological models describe people's interactions with physical and social environments on the basis of the theory that human behaviors are promoted or hindered by environments (Wicker, 1984; Sallis, Bauman, & Pratt, 1998). Multiple levels of environment systems and different categories of environmental influence were built upon the research by Bronfenbrenner (1977) and McLeroy et al. (1988).

Bronfenbrenner (1977) proposed three levels of hierarchy — micro-, meso-, and exo-system — to describe human behavior and its interaction with different levels of surrounding settings. McLeroy et al. (1988) suggested five different categories for influential factors, including intra- and inter-personal, community, institutional, and policy factors.

Building upon the theoretical base of ecological models, many scholars have put much effort into improving them. King et al. (2002) proposed the application of the perspectives in different fields, including social ecology and urban planning. The urban planning perspective was highlighted with the categorization of urban features according to pedestrian-friendly and vehicle-oriented environments. Spence and Lee (2003) expanded the ecological system theory of Bronfenbrenner (1977) by adding four more aspects, including biological and psychological factors, which may be closely related to PA behavior. Sallis, Bauman, and Pratt (1998) focused on supportive policies for PA promotion, apart from the environmental aspect of the ecological model. The policy was then incorporated in the more comprehensive ecological model by Sallis et al. (2012) with other factors, including individual characteristics, social/cultural, and built-environment factors. Common social factors which showed empirical evidence or have theoretical associations with PA, including social support and social cohesion, were also incorporated in the ecological model suggested by McNeill, Kreuter, and Subramanian (2006).

In practice, studies to date have sought empirical evidence on the relationships of individuals' PA with built environments, safety, aesthetics, and social environments in neighborhood areas, controlling individual characteristics. Especially for adults, objectively measured population densities, land-use mix, and street length/connectivity had strong associations with PA (Troped et al., 2010). Using GPS and accelerometer devices to objectively

measure the movements and PA levels of adults, Boruff, Nathan, and Nijënstein (2012) found positive associations between PA and recreational and park areas.

However, the findings of Hoehner et al. (2005) indicated that there was no significant association between adults' transportation or recreational PA and places to exercise, including nearby parks, trails, and private fitness facilities. Residential and commercial areas, transit stops/lines, and sidewalks also showed inconsistent associations (Hoehner et al., 2005; Nagel et al., 2008; Troped et al., 2010; Boruff, Nathan, & Nijënstein, 2012). Institutional areas, including schools and community services, and walkability did not show any significant associations with adults' PA. Objectively measured traffic volume had a positive association with adults' walking time, whereas the low traffic volume showed mixed associations with their leisure time PA in different regions (McGinn et al., 2007; Nagel et al., 2008). Regarding social and physical environmental factors, existing studies failed to produce consistent findings to provide empirical evidence for ecological models. Therefore, this study seeks to bridge the gap between what the underpinning theories suggest regarding the interactions and mixed findings of studies in practice by considering complex interactions between ATMs and various social and physical environmental factors considering daily paths of individuals.

1.3.2 Delineation of dynamic exposure areas to environmental contexts

Before tracking devices combined with an accelerometer sensor were widely used in PA research to consider daily movements of individuals, a number of scholars attempted to find influential factors in the proximate environments around individual home locations, since people usually spend much time visiting places around their homes. For example, some studies investigated the impact of green spaces in residential neighborhoods on PA using GIS techniques (Cohen et al., 2006; Coombes, Jones, & Hillsdon, 2010; Nagel et al., 2008; Schipperijn et al.,

2013). The levels of PA are identified by perceived measures through survey or objective measures using an accelerometer device. Residential neighborhoods are delineated by circular areas around each home location using buffer analysis in GIS, and characteristics of environmental contexts within these buffer areas (e.g., access to green spaces) are assessed to statistically examine the associations with the measures of PA. In terms of neighborhood delineation using buffer analysis, previous studies with different types (e.g., circular, network) and sizes (e.g., 1000m, 1600m, 3000m) were especially reviewed to find reliable buffer parameters showing that greenness best predicts physical health (Browning & Lee, 2017).

GPS caused a great change in PA research as its use allowed researchers to look further into daily PA levels tied to places people visit and time that they spend there. GPS trajectories with GIS data and analysis collectively worked to resolve such uncertainties in space and time, which raised UCGoP. With regard to the GIS analysis, previous studies mostly used circular buffers generated along individuals' space-time paths to estimate dynamic exposures to green spaces, land use including commercial places, and safety (Rodríguez et al., 2012; Harrison et al., 2014; Burgoine et al., 2015).

Due to a lack of consensus on buffer characteristics in the existing studies, research findings were, however, inconsistent regarding the associations between PA and environmental factors. The buffer distance, especially, is one of the most influential settings that may cause the research outcomes. Thus, the varying contextual influences need to be further investigated, taking into account different buffer distance settings to provide insight into understanding the implication of the buffer distance on the association between PA and each environmental factor, and to suggest a reliable buffer distance, resulting in less inconsistency in the interpretation of the impact of environmental factors and greater statistical significance of the factors.

1.3.3 Machine learning models in PA research

In PA and transportation research, machine learning has been widely used for the classification of activity types (PA types or travel modes), using data collected from some sensor(s). As a branch of artificial intelligence, machine learning involves a training process to predict outcomes based on machine learning models/algorithms, using a large quantity of input data. Quantifiable values called 'features' are derived from the input data to train machine learning models. Machine learning techniques have helped in accurate recognition of walking, jogging, standing, biking, car, bus, etc. using time-series data from one or more sensors, including GPS and accelerometer (Kwapisz, Weiss, & Moore, 2011; Arif et al., 2014; Ellis et al., 2014; Fang et al., 2016). The identification of different activity types is primarily useful for encouraging people to engage in PA by providing feedback through mobile applications, and determining abnormal behaviors when monitoring elderly people for their healthcare.

Besides the classification of activity types, machine learning has the potential to provide insights into complex interactions between human behaviors and environmental contexts, which previous studies in PA research have not yet explored. Contextual influences in ecological models of health behavior, like travel modes, have multiple levels including demographics, physical and social environments, and community. One promising research perspective in the PA field is to not only enhance the understanding of various factors associated with PA but also the interactions between factors across different levels (Sallis, Owen, & Fisher, 2015). Machine learning can be very helpful in facilitating a thorough understanding of complex interactions, since it has recently evolved to interpret models for accurate explanations of the influence of various features on specific outcomes (Letham et al., 2015; Krause et al., 2016). In other words, machine learning techniques can facilitate a more in-depth understanding of travel modes and

environmental contexts. Thus, this study seeks to provide a novel approach to the interpretation of the influence of various environmental factors on the predictions of ATMs, using machine learning techniques.

1.4 DISSERTATION ORGANIZATION

This dissertation discusses various methodological aspects that will be required to advance health and transportation research. In the next chapter, an automatic travel mode classification algorithm using GPS and accelerometer data is described. Chapter 3 delves into the existence of buffer-size effects in statistical analyses, which may affect research findings of the associations between travel modes and environmental factors. In Chapter 4, a novel framework to find meaningful patterns of the associations between travel modes and environmental contexts taking into account a large amount of movement data of individuals with the aid of machine learning techniques, and interpret the findings using explanatory tools. Chapter 5 summarizes main research findings, contributions, implications, and future work.

1.5 REFERENCES

Almanza, E., Jerrett, M., Dunton, G., Seto, E. & Pentz, M. A. (2012). A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. *Health & place* **18** (1), 46–54.

Arif, M., Bilal, M., Kattan, A., Ahamed, S. I. (2014). Better physical activity classification using smartphone acceleration sensor. *Journal of Medical Systems*, 8, 95.

Boruff, B. J., Nathan, A. & Nijлnstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood'in place-based health research. *International journal of health geographics*, 11, 22.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American psychologist* **32** (7), 513.

Browning, M. & Lee, K. (2017). Within what distance does "Greenness" best predict physical health? A systematic review of articles with GIS buffer analyses across the lifespan. *International journal of environmental research and public health* **14** (7), 675.

Burgoine, T., Jones, A. P., Brouwer, R. J. N. & Neelon, S. E. B. (2015). Associations between BMI and home, school and route environmental exposures estimated using GPS and GIS: do we see evidence of selective daily mobility bias in children?. *International journal of health geographics* **14** (1), 8.

Cohen, D. A., Ashwood, J. S., Scott, M. M., Overton, A., Evenson, K. R., Staten, L. K., Porter, D., McKenzie, T. L. & Catellier, D. (2006). Public parks and physical activity among adolescent girls. *Pediatrics* **118** (5), e1381–e1389.

Coombes, E., Jones, A. P. & Hillsdon, M. (2010). The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Social science &*

medicine **70** (6), 816–822.

Diez Roux, A. V., Mair, C., 2010. Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186, 125–145.

Duncan, S., White, K., Mavoa, S., Stewart, T., Hinckson, E. & Schofield, G. (2016). Active transport, physical activity, and distance between home and school in children and adolescents. *Journal of Physical Activity and Health* **13** (4), 447–453.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J. & Kerr, J. (2014). Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Public Health*, 2, 39–46.

Fang, S.-H., Liao, H.-H., Fei, Y.-X., Chen, K.-H., Huang, J.-W., Lu, Y.-D. & Tsao, Y. (2016). Transportation modes classification using sensors on smartphones. *Sensors*, 16, 1324.

Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public health nutrition* **2** (3a), 411–418.

Gordon-Larsen, P., Nelson, M. C. & Beam, K. (2005). Associations among active transportation, physical activity, and weight status in young adults. *Obesity Research* **13** (5), 868–875.

Gordon-Larsen, P., Nelson, M. C., Page, P. & Popkin, B. M. (2006). Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics* **117** (2), 417–424.

Harrison, F., Burgoine, T., Corder, K., van Sluijs, E. M. & Jones, A. (2014). How well do modelled routes to school record the environments children are exposed to?: a cross-sectional comparison of GIS-modelled and GPS-measured routes to school. *International journal of health geographics* **13** (1), 5.

Helbich, M., van Emmichoven, M. J. Z., Dijst, M. J., Kwan, M.-P., Pierik, F. H. & de Vries, S. I.

(2016). Natural and built environmental exposures on children's active school travel: A Dutch global positioning system-based cross-sectional study. *Health & place* **39**, 101–109.

Hoehner, C. M., Ramirez, L. K. B., Elliott, M. B., Handy, S. L. & Brownson, R. C. (2005). Perceived and objective environmental measures and physical activity among urban adults. *American journal of preventive medicine* **28** (2), 105–116.

Jansen, M., Ettema, D., Pierik, F. & Dijst, M. (2016). Sports facilities, shopping centers or homes: What locations are important for adults' physical activity? A cross-sectional study. *International journal of environmental research and public health* **13** (3), 287.

King, A. C., Stokols, D., Talen, E., Brassington, G. S. & Killingsworth, R. (2002). Theoretical approaches to the promotion of physical activity. *American journal of preventive medicine* **23** (2), 15–25.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697). ACM.

Kwan, M.-P. (2012a). How GIS can help address the uncertain geographic context problem in social science research. *Annals of GIS* **18** (4), 245–255.

Kwan, M.-P. (2012b). The uncertain geographic context problem. *Annals of the Association of American Geographers* **102** (5), 958–968.

Kwan, M.-P. (2013). Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers*, 103, 1078–1086.

Kwan, M.-P. (2018a). The limits of the neighborhood effect: Contextual uncertainties in

geographic, environmental health, and social science research. *Annals of the American Association of Geographers*, **108** (6), 1482–1490.

Kwan, M.-Po. (2018b). The neighborhood effect averaging problem (NEAP): An elusive confounder of the neighborhood effect. *International Journal of Environmental Research and Public Health*, **15,** 1841.

Kwapisz, J. R., Weiss, G. M. & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12, 74–82.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, **9** (3), 1350–1371.

McGinn, A. P., Evenson, K. R., Herring, A. H., Huston, S. L. & Rodriguez, D. A. (2007). Exploring associations between physical activity and perceived and objective measures of the built environment. *Journal of Urban Health* **84** (2), 162–184.

McLeroy, K. R., Bibeau, D., Steckler, A. & Glanz, K. (1988). An ecological perspective on health promotion programs. *Health education quarterly* **15** (4), 351–377.

McNeill, L. H., Kreuter, M. W. & Subramanian, S. (2006). Social environment and physical activity: a review of concepts and evidence. *Social science & medicine* **63** (4), 1011–1022.

Nagel, C. L., Carlson, N. E., Bosworth, M. & Michael, Y. L. (2008). The relation between neighborhood built environment and walking activity among older adults. *American journal of epidemiology* **168** (4), 461–468.

Oliver, M., Mavoa, S., Badland, H., Parker, K., Donovan, P., Kearns, R. A., Lin, E.-Y. & Witten, K. (2015). Associations between the neighbourhood built environment and out of school

physical activity and active travel: an examination from the Kids in the City study. *Health & place* **36**, 57–64.

Perchoux, C., Chaix, B., Cummins, S. & Kestens, Y. (2013). Conceptualization and measurement of environmental exposure in epidemiology: accounting for activity space related to daily mobility. *Health & Place*, 21, 86–93.

Physical Activities Guidelines Advisory Committee. (2008). Physical activity guidelines advisory committee report. Washington (DC): US Department of Health and Human Services.

Rodriguez, D. A., Cho, G.-H., Evenson, K. R., Conway, T. L., Cohen, D., Ghosh-Dastidar, B., Pickrel, J. L., Veblen-Mortenson, S. & Lytle, L. A. (2012). Out and about: association of the built environment with physical activity behaviors of adolescent females. *Health & place* **18** (1), 55–62.

Sallis, J., Bauman, A. & Pratt, M. (1998). Environmental and policy interventions to promote physical activity. *American journal of preventive medicine* **15** (4), 379–397.

Sallis, J. F., Cervero, R. B., Ascher, W., Henderson, K. A., Kraft, M. K., & Kerr, J. (2006). An ecological approach to creating active living communities. *Annual Review of Public Health*, **27** (1), 297–322.

Sallis, J. F., Floyd, M. F., Rodriguez, D. A. & Saelens, B. E. (2012). Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* **125** (5), 729–737.

Sallis, J. F., Owen, N. & Fisher, E. (2015). Ecological models of health behavior. *Health behavior: theory, research, and practice. 5th ed. San Francisco: Jossey-Bass*, 43–64.

Schipperijn, J., Bentsen, P., Troelsen, J., Toftager, M. & Stigsdotter, U. K. (2013). Associations between physical activity and characteristics of urban green space. *Urban Forestry &*

Urban Greening **12** (1), 109–116.

Spence, J. C. & Lee, R. E. (2003). Toward a comprehensive model of physical activity. *Psychology of sport and exercise* **4** (1), 7–24.

Troped, P. J., Wilson, J. S., Matthews, C. E., Cromley, E. K. & Melly, S. J. (2010). The built environment and location-based physical activity. *American journal of preventive medicine* **38** (4), 429–438.

Voss, C., Sims-Gould, J., Ashe, M. C., McKay, H. A., Pugh, C. & Winters, M. (2016). Public transit use and physical activity in community-dwelling older adults: Combining GPS and accelerometry to assess transportation-related physical activity. *Journal of Transport & Health* **3** (2), 191–199.

Wei, M., Gibbons, L. W., Kampert, J. B., Nichaman, M. Z. & Blair, S. N. (2000). Low cardiorespiratory fitness and physical inactivity as predictors of mortality in men with type 2 diabetes. *Annals of internal medicine* **132** (8), 605–611.

Wicker, A. W. (1984). *An introduction to ecological psychology.* CUP Archive.

World Health Organization. (2011). Global recommendations on physical activity for health. Retrieved from https://www.who.int/dietphysicalactivity/physical-activity-recommendations-18-64years.pdf

CHAPTER 2: AUTOMATIC TRAVEL MODE CLASSIFICATION

2.1 INTRODUCTION

A large quantity of movement data has been collected and analyzed in various research domains since tracking technology was advanced (Eagle et al., 2009; Gonzalez et al., 2008; Shoval and Isaacson, 2007). As one of the tracking devices, GPS receivers have been widely used to collect data that enhance our understanding of the spatiotemporal dynamics of moving objects, such as animals (Dodge et al., 2013; Laube et al., 2007), humans (Kwan, 2004; Shoval et al., 2011; Wang et al., 2018), or vehicles (Downs and Horner, 2012; Ferreira et al., 2013). In PA research, GPS trajectories and the capabilities of GIS facilitate a better understanding of the associations between moderate to vigorous physical activity (PA), such as brisk walking and running, and various environmental factors, which takes individuals' daily travels into account (Almanza et al., 2012; Boruff et al., 2012; Cooper et al., 2010; Jansen et al., 2016; Lachowycz et al., 2012; Rodríguez et al., 2012; Troped et al., 2010). Such recent PA studies that use GPS trajectories revealed the importance of non-residential areas on people's health behaviors and outcomes in addition to residential neighborhoods (Diez Roux and Mair, 2010; Perchoux et al., 2013). High-resolution GPS data can also be used to mitigate the uncertain geographic context problem (UGCoP), since they help identify the various non-residential contexts that may affect people's health (Kwan, 2012a,b; 2013).

PA research has largely used objectively measured accelerometer data to assess the intensity of people's PA, since objective PA measures yield more significant findings than subjective and self-reported measures (Browning and Lee, 2017). PA research, however, needs to widen its focus from the intensity of PA to types of PA, which may provide useful clues for

understanding specific health behaviors in particular geographic contexts (Jankowska et al., 2015). Some PA conceptual models suggest that specific types of travel modes, like walking and biking, have associations with specific environmental or geographic contexts, such as trails, safe pedestrian sidewalks, and supportive facilities. Yet, more consistent empirical evidence is needed to support such associations (Loukaitou-Sideris, 2006; Sallis et al., 1998). In addition, few PA studies to date using GPS and accelerometer data have taken into account inactive travel modes, like traveling in a vehicle (e.g., Voss et al., 2016). According to Gordon-Larsen et al. (2005), the high utilization of active travel modes (ATMs) (e.g., walking and biking) for commuting is mostly exhibited among the group of young adults who meet PA recommendations, whereas higher percentage of young adults who do not meet PA recommendations is associated with higher utilization of inactive travel models (e.g., public transit and private car) for traveling to workplaces or schools. Further, the impact of various environmental factors on people's PA may also be less when they stay inside a vehicle, compared to when they walk or run outside. It is thus important to identify and separate motorized transport modes (vehicle-based movement) from PA (non-vehicle-based human movement) in order to more accurately estimate people's exposure to various environmental contexts. However, only a few PA studies have attempted to identify whether people are traveling in vehicles (in-vehicle status) or not through the automatic classification of travel modes to date (Ellis et al., 2014; Zhou, 2014).

To contribute to this literature, this study proposes and develops an algorithm to automatically classify travel modes using GPS and accelerometer data available to the public. Hierarchical classification processes based on machine learning techniques are innovative approaches adopted in this study. As a branch of artificial intelligence, machine learning improves the prediction of outcomes through a training process based on machine learning

models/algorithms using a large amount of input data. The introduction of hierarchical classification is to jointly identify more classes (identified with distinctive labels) — like the different types of PA in this study — using heterogeneous sensor data, such as GPS and accelerometer data, than those using only one of these two. The hierarchical classification processes are discussed in detail in Section 2.3. In this study, three components constitute the framework of the hierarchical classification algorithm: indoor/outdoor classification, classification using GPS data (outdoors), and classification using accelerometer data (mostly indoors). Machine learning techniques make predictions of people's travel modes based on the generated numeric or categorical features from collected GPS and accelerometer data. As measurable characteristics, features are informative quantifiable properties calculated from input data and one of the fundamental elements in machine learning used to predict different classes (e.g., travel modes in this study). Regarding predicted classes, biking, running, walking, standing, sitting (sedentary status), and traveling in a vehicle (in-vehicle status) are automatically identified through the developed algorithm. Because running and biking are the two most popular outdoor PA types for young adults in the U.S. (Outdoor Foundation, 2015), the proposed algorithm will seek to identify these two travel modes.

The proposed algorithm is quantitatively and qualitatively validated using real-world GPS data collected from three subjects in highly and moderately urbanized areas. Highly urbanized areas here mean areas with high building density, tall buildings and heavy traffic, while moderately urbanized areas are areas with moderate building density, fewer tall buildings and moderate traffic. The rationale for choosing highly- and moderately urbanized areas in this study is that the degree of urbanization (e.g., tall buildings and traffic congestion) may influence the accuracy of GPS measurement. For example, in highly urbanized areas, GPS positional errors

are likely to be higher than those in moderately urbanized areas due to the obstruction of GPS signals by tall buildings (known as urban canyons in GPS parlance). Further, travel speed tends to be very slow in certain highly congested road segments in highly urbanized areas, which may affect classification accuracy of in-vehicle status since the GPS position of a moving object becomes unstable when it is stationary or moving slowly. Therefore, the performance of the algorithm needs to be explored in the two areas with different levels of urbanization.

In this study, Chicago was chosen as an example of highly urbanized areas, and Champaign County in Illinois was selected as an example of moderately urbanized areas. Chicago is the second largest city in the U.S. with the most skyscrapers that are 490 feet (150 m) or greater in height whereas Champaign has no skyscraper (Council on Tall Buildings and Urban Habitat, 2018). Chicago also ranked third in total travel delay and total congestion cost among the largest urban areas in the U.S. in 2014 (Schrank et al., 2015). Regarding total travel delay, Champaign has 1,966,000 hours of extra travel time in 2014, which is much less than Chicago (302,609,000 hours). As home to the University of Illinois at Urbana-Champaign, Champaign has a population of 201,081, and is the 10th-most populous county in Illinois with diverse educational, cultural, and recreational opportunities for local residents (United States Census Bureau, 2010). When compared to Champaign County, the landscape of Chicago is mostly comprised of high-rise buildings, and it has relatively a high level of traffic congestion. This is, in turn, expected to be more challenging environment for predictions of travel modes because of GPS signal obstruction by tall buildings and stagnant GPS points of in-vehicle status by traffic congestions, which are likely to be identified as relatively slow travels, like biking.

The classification algorithm can greatly reduce participants' burdens in recording travel modes by automatically classifying them and is thus useful for human mobility research using

GPS, including transportation and public health studies. Researchers can also benefit from the automatic classification since there is no need to be concerned about missing or inaccurate travel mode records made by participants. Automatic travel mode detection will also be useful for identifying travel patterns associated with particular route choices and for understanding the underlying decision-making processes in route choice research (Broach et al., 2012; Papinski et al., 2009). In addition, this study shows how published heterogeneous sensor data — GPS and accelerometer data — can work together to provide accurate and automatic PA classification using machine learning and geospatial techniques.

The rest of Chapter 2 is organized into four sections. Section 2.2 describes previous studies on travel mode classification using both GPS and accelerometer data. Section 2.3 describes the algorithm with respect to its components, focusing more on indoor/outdoor classification and classification using GPS data. Section 2.4 validates the proposed algorithm, and Section 2.5 discusses the research findings and provides conclusions.

2.2 PAST STUDIES ON TRAVEL MODE CLASSIFICATION

Many PA recognition studies have been conducted, leveraging the potential of the sensing capabilities of smartphones with a particular focus on health promotion and management. Particularly, the accelerometer sensors commonly equipped in most smartphones brought profound enhancement to the automatic recognition of travel modes, including running, walking, and sitting with high precision using machine learning techniques (Anguita et al., 2012; Arif et al., 2014; Kwapisz et al., 2011; Weiss et al., 2016; Zhang et al., 2010). For instance, travel mode classification was employed in a web-based application to monitor the PA of

children, obese people, or the elderly and to encourage them to perform sufficient PA in their daily lives (Weiss et al., 2016).

The automatic recognition of different travel modes using GPS trajectories has drawn the attention of transport researchers due to the low response rates and high incompletion rates of paper- or phone-based travel surveys. For instance, Zheng et al. (2010) proposed an approach to classify four kinds of travel modes – walking, driving, traveling by bus, and biking – by segmenting GPS trajectories and extracting features like maximum velocity and acceleration in each GPS segment for understanding human mobility and displaying it on web-based mapping applications (Table 2.1). The study suggests a way of dividing a GPS trajectory into walking and non-walking segments stressing the importance of the walking mode as a transition to a non-walking mode like car, bus, or train. With the segmentation approach and the use of a decision tree, the study achieved a predictive accuracy of over 75%. Along with GPS data, GIS data were employed for achieving higher accuracy in travel mode classification. In most cases, transport network data and data of related infrastructure (e.g., stops, stations, and entrances) have been used (Biljecki et al., 2013; Gong et al., 2012; Witayangkurn et al., 2013). For example, Biljecki et al. (2013) extracted road, railway, bus, and tram networks, the locations of bus stops, and train stations from OpenStreetMap to detect ten transport modes using a fuzzy logic approach, and the algorithm achieved 92% accuracy. Besides transport links, Moiseeva et al. (2010) proposed a travel mode inference system for the classification of 7 transport modes using land use data. The land use data helped increase the predictive accuracy for particular kinds of transport modes that may occur on certain types of land use (e.g., railroad tracks). The transport mode classification using a Bayesian belief network model achieved a 95% accuracy. On the other hand, some studies did not use GIS data because these data might not be available for many study areas.

However, some of these studies were able to achieve moderate to high predictive accuracy (Xiao et al., 2017; Zhu et al., 2016).

With respect to performance improvement, how features are extracted can significantly affect the results of travel mode classification. For the last decade, some studies considered the focal characteristics of sub-segments for each trip captured through moving windows sliding on GPS trajectories (Bolbol et al., 2012; van Dijk, 2018; Dodge et al., 2009; Xiao et al., 2017). A moving window is a calculation method to analyze aggregated characteristics of data points by computing a series of derivatives (e.g., average speed) from different subsets. Bolbol et al. (2012) applied a fixed-size moving window sliding on speed and acceleration values of multi-segment GPS instances, and the classification of six travel modes using support vector machine (SVM) achieved an accuracy of 88%. Dodge et al., (2009) and Xiao et al. (2017) especially adopted the focal characteristic of GPS trajectories by calculating movement parameters from the GPS points that fall within a sliding window, which achieved a predictive accuracy of 82% and 91% respectively. van Dijk (2018) achieved over 99% predictive accuracy in classifying trips (moving) and activities (stay) by introducing moving spatial and temporal windows.

The combined use of multiple sensor data has recently been introduced to travel mode classification research due to the increasing availability of various sensors. Patrick et al., (2008) developed the Physical Activity and Location Measurement System (PALMS) to understand PA-related energy expenditure associated with time and space in exposure biology research by incorporating GPS, accelerometer and heart rate monitoring sensors. With regard to travel mode identification, accelerometer, magnetometer, and gyroscope data recorded using smartphones enhanced the performance of travel mode detection (Ellis et al., 2014; Fang et al., 2016; Feng and Timmermans, 2013; Shafique and Hato, 2016; Zhou, 2014). Ellis et al. (2014) particularly

compared the predictive accuracy of several machine learning models for the classification of six travel modes, including bus, car, sitting, and walking, based on GPS and accelerometer data collected from two trained assistants in different built environment settings. Evenson and Furberg (2017) developed a smartphone application for users and researchers to automatically predict PA types using GPS and accelerometer data.

Table 2.1. Characteristic properties of past studies on travel modes

Author	Predictive accuracy	Classification algorithm or system	Sensor	Classified activities	Observation unit	GIS data use	Moving window
Zheng et al. (2010)	75.60%	Tree-based model	GPS	walking, biking, car, bus,	Segment	No	No
Biljecki et al. (2013)	91.60%	Fuzzy expert system	GPS	walking, biking, car, bus, train, tram, underground, ferry, sailing boat, aircraft	Segment	Yes	No
Moiseeva et al. (2010)	95.40%	Bayesian Belief Network	GPS	walking, running, biking, motorbike, car, bus, train	Segment	Yes	No
Xiao et al. (2017)	90.77%	XGB	GPS	walking, biking, car, bus and taxi, subway, train	Segment	No	Yes
Zhu et al. (2016)	91.44%	RF	GPS	walking, biking, car, bus	Segment	Yes	No
Bolbol et al. (2012)	88.00%	SVM	GPS	walking, biking, car, bus, train, underground,	Segment	No	Yes
Dodge et al. (2009)	82.00%	SVM	GPS	pedestrian, biking, car, motorbike	Segment	No	Yes
van Dijk (2018)	99.40%	SVM and RF	GPS	move, stay	Point	Yes	Yes

Table 2.1 (cont.)

Author	Predictive accuracy	Classification algorithm or system	Sensor	Classified activities	Observation unit	GIS data use	Moving window
Ellis et al. (2014)	91.90%	RF	GPS, accelerometer	walking, biking, car, bus, sitting, standing	Point	No	Yes
Fang et al. (2016)	86.94%	SVM	Accelerometer, magnetometer, gyroscope	walking, running, biking, vehicle, stay	Segment	No	Yes
Feng and Timmermans (2013)	85%	Bayesian Belief Network	GPS, accelerometer	walking, running, biking, motorbike, car, bus, tram, subway	Point	No	Yes
Shafique and Hato (2016)	99.96%	RF	Accelerometer	walking, biking, car, bus, train, subway	Segment	No	Yes
Zhou (2014)	Over 80%	RF	GPS and accelerometer	walking, running, biking, in-vehicle, stay	Point	No	No

* RF: random forest, SVM: support vector machine, XGB: extreme gradient boosting

In this study, an automatic travel mode classification algorithm is developed using two kinds of open and publicly available sensor data — GPS and accelerometer data — collected by other researchers (these data will be described in Section 2.3.1 below). Being sedentary is an important stationary behavior, which many PA studies also examined in addition to moderate to vigorous PA. The contribution of this study lies on how to utilize two different kinds of sensor data to identify travel modes. As adopted in the research by Dodge et al., (2009) and Xiao et al. (2017), the merit of the GPS trajectory operators at the focal level used in this study is that they can help capture distinctive variations in movement derivatives (e.g., velocity, acceleration) over time for different travel modes. This study, in particular, extensively uses GPS trajectory

operators at the focal level with different sizes of moving spatial and temporal windows, which are described in Section 2.3.2. Compared to the research by van Dijk (2018) that used spatial and temporal moving windows (as shown in Table 2.1), this study extracts more features to maximize the benefit of spatial and temporal moving windows for classifying more travel modes by utilizing two different kinds of open sensor data.

2.3 A FRAMEWORK OF AUTOMATIC CLASSIFICATION OF TRAVEL MODES USING PUBLICLY AVAILABLE GPS AND ACCELEROMETER DATA

The automatic travel mode classification in this study is implemented through hierarchical classification processes to jointly identify a total of six classes — biking, running, walking, standing, being sedentary, and riding in a vehicle (Figures 2.1 and 2.2). Hierarchical classification has been used to deal with hierarchical structures in real-world systems regarding classification from the top level to lower levels (Dumais and Chen, 2000; McNamara et al., 2015). In this study, a hierarchical classification approach is expected to greatly facilitate the accurate classification of indoor versus outdoor activities at the top level, and vehicle-based versus non-vehicle-based travel modes using GPS and accelerometer data at the lower levels (Figure 2.1). That is because GPS data are suitable for identifying vehicle-based movement versus non-vehicle-based human movement, whereas accelerometer data are commonly used for classifying various types of non-vehicle-based human movement. In particular, some travel modes (e.g., standing and sitting) cannot be characterized by using GPS data but can be identified by using accelerometer data. A hierarchical classification approach thus allows for the identification of these travel modes among non-vehicle-based movements at a lower level using accelerometer data (Figure 2.1). Apart from these two classification processes, GPS points are

also classified into indoor or outdoor points at the top level to facilitate the accurate classification processes at lower levels. For example, in an indoor environment, basic types of PA, like walking and standing, are likely to occur, whereas in-vehicle status and biking are unlikely to take place indoors. Therefore, indoor GPS points should go through a classification process that involves the analysis of accelerometer data. In this study, indoor GPS points and accelerometer data are processed at the lower level through the 'Classification using accelerometer data' process for predicting specific indoor travel modes. To classify the indoor or outdoor context, I used measures solely derived from the GPS data not involving any additional sensor data (e.g., camera images, light sensor levels) to enhance the performance of the methods implemented in previous studies (Lam et al., 2013; Tandon et al., 2013).

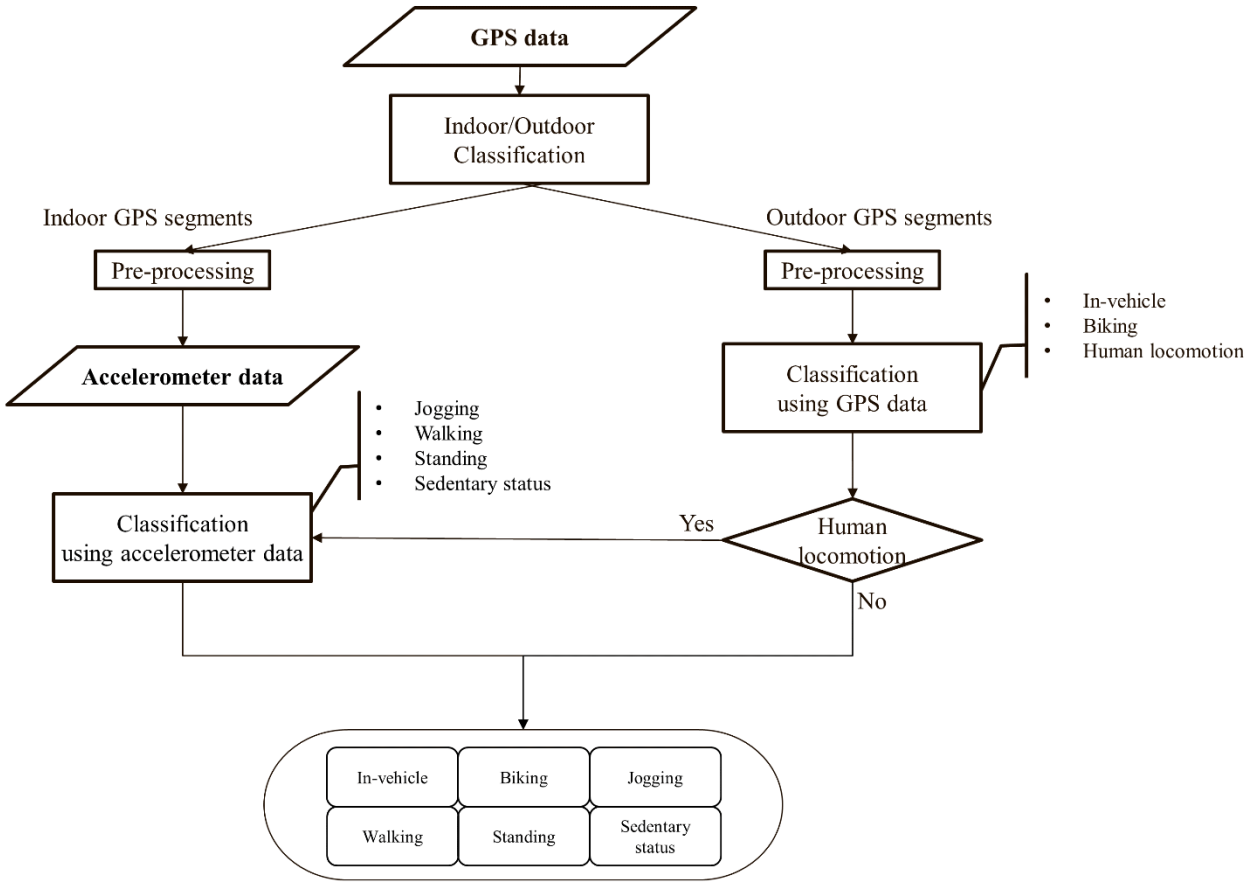


Figure 2.1. Flow diagram of travel mode classification algorithm

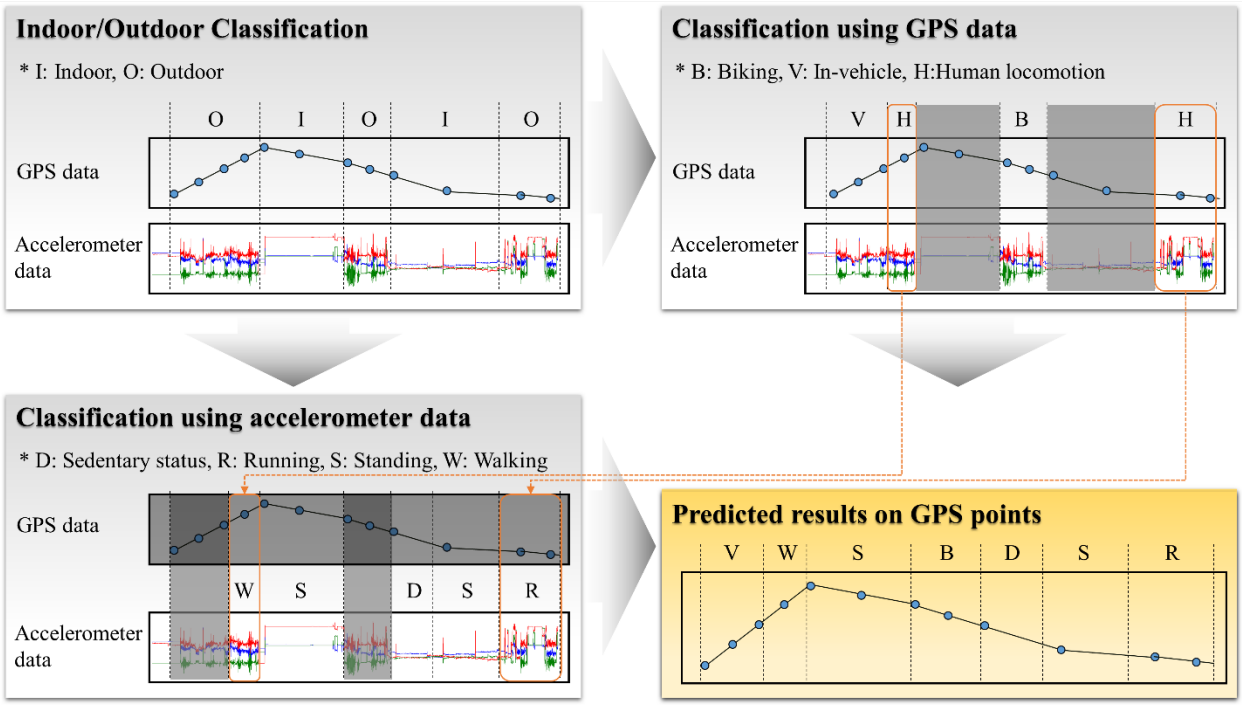


Figure 2.2. Details of the three classification components (Indoor/Outdoor classification, classification using GPS data, and classification using accelerometer data) and hierarchical classification of human locomotion

Among the three-level classification processes shown in Figures 2.1 and 2.2, the 'Indoor/Outdoor Classification' process first partitions GPS trajectories into indoor or outdoor segments since indoor GPS segments are ineffective and unreliable for extracting any movement derivatives due to the low accuracy and loss of GPS signs in indoor environments. Outdoor GPS segments identified in this process will then go through the 'Classification using GPS data' process, which classifies GPS segments into vehicle-based movement (riding in a vehicle and biking) and non-vehicle-based human movement (human locomotion). The indoor/outdoor classification and classification using GPS data processes apply the concept of GPS trajectory operators (Laube et al., 2007) discussed in Section 2.3.2. On the other hand, indoor GPS

segments are not used to extract features for any classification process but used to provide time ranges within which specific indoor activities like sedentary status and walking are predicted through ‘Classification using accelerometer data’ (Lee and Kwan, 2018). The pre-processing is followed by the classification using GPS or accelerometer data to filter out low-quality GPS points. Among the three identified classes in the ‘Classification using GPS data’ process, human locomotion is further classified into specific travel modes such as running and walking through the “Classification using accelerometer data’ process. Through these three-level classification processes, six different classes are identified: traveling in a vehicle, biking, running, walking, standing, and being sedentary.

2.3.1 Data description

Datasets from four different sources were used for the three-level classification processes. First, since there are no publicly available GPS data that can be used to classify indoor versus outdoor trajectories, GPS data were collected from three persons living in highly or moderately urbanized areas for 7 days with horizontal dilution of precision (HDOP) and ‘indoor’ or ‘outdoor’ labels. Dilution of precision represents the precision of positional measurement, and specifically, HDOP indicates the geometric quality of the horizontal positions of GPS data. A large discrepancy in GPS accuracy between highly and moderately urbanized areas might affect the indoor/outdoor identification. Thus, GPS and accelerometer data were collected from one subject living in a highly urbanized area and two subjects who live in a moderately urbanized area. These data were used in the test phase reported in Section 2.4.1 below. Only two days of the GPS trajectories, which have many combinations of indoor and outdoor trips, from each of the three subjects were also used in the indoor/outdoor classification in Section 2.3.3.

Second, version 1.3 of the Geolife project GPS dataset (Microsoft Research Asia) is used for the classification using GPS data. The GPS dataset was collected from 182 subjects from April 2007 to August 2012 (Zheng et al., 2008; Zheng et al., 2010). Most of these GPS trajectories were tracked at short time intervals, like every 1 to 5 seconds, in Beijing, China. Since it is publicly available data, no demographic information regarding participants was available. Only a part of the GPS data from 73 participants have labels of 11 kinds of transport modes. Among them, GPS trajectories labeled with 7 transport modes – train, bus, car, taxi, biking, walking, and running – are used in the study. The total number of GPS points with these 7 transport-mode labels is 5,063,475, and walking (35%) comprises a large proportion of the GPS points whereas running (0.03%) constitutes the smallest portion.

Third, due to the lack of GPS data for the running mode, open GPS data recorded when people were running were obtained from OpenStreetMap and TrackProfiler and used in this study. Train, taxi, bus, and car transport modes are merged into one motorized mode, and walking and running (GPS data from Geolife, OpenStreetMap, and TrackProfiler) are grouped into one human locomotion class in the study.

Finally, version 1.1 of the Wireless sensor data mining accelerometer dataset is used for the classification using accelerometer data process (Kwapisz et al., 2011). Three-axis acceleration was collected from 36 persons with timestamp under laboratory conditions and recorded at 20 Hz. Each instance (observation) has a travel mode label like walking, running, or standing. Since movements on stairs account for a small portion of the daily PA of the subjects, movements upstairs and downstairs are merged into walking in this study.

Table 2.2. Description of GPS and accelerometer datasets

Dataset	Number of instances	Attributes	Mode	Sampling rate
GPS dataset collected from two subjects	58,172	Person ID, longitude, latitude, time stamp, HDOP, and indoor/outdoor mode	Indoor (61%), outdoor (39%)	1 second
Geolife GPS dataset v1.3 (Microsoft Research Asia)	5,063,475	Person ID, longitude, latitude, time stamp, and transport mode	Train, (5%), Bus (17%), car (8%), taxi (7%), bike (28%), walk (35%), and run (0.03%)	Mostly 1~5 seconds
GPS dataset from OpenStreetMap and TrackProfiler	10,435	Longitude, latitude, time stamp	Run (100%)	Variable - 1~17 seconds
Wireless sensor data mining's accelerometer dataset v1.1	1,098,207	Person ID, physical activity mode, timestamp, and acceleration (x, y, z)	Walking (39%), Running (31%), Upstairs (11%), Downstairs (9%), Sitting (6%), and Standing (4%)	20hz (20 samples / sec)

2.3.2 GPS trajectory operators

GPS trajectory operators (Laube et al., 2007) are used to extract features from GPS trajectories in the indoor/outdoor classification and classification using GPS data processes. Different levels of operators for GPS trajectories were proposed by introducing map algebra operations to capture dynamic movement characteristics over space and time. An operator performs specific mathematical and/or logical analysis or calculation. For instance, in map algebra, there are four types of operations, namely local, focal, zonal, and global, for spatial analysis to produce a resulting map using raster data (Tomlin, 1990). Among the four different levels of GPS operators (instantaneous, interval, episodal, and global), instantaneous and interval operators are used to acquire the local and focal characteristics of dynamic movements and to calculate movement derivatives (speed, acceleration) for feature extraction. In machine learning, feature extraction is a process for deriving values (features) from the data to train a model. Movement derivatives at the instantaneous level are calculated between two consecutive GPS points to capture local movement characteristics pertinent to each of these GPS points. For

instance, assuming that a given GPS trajectory is recorded at a two-second interval (as shown in Figure 2.3), instantaneous velocity of the GPS point p_3 is calculated based on the distance and two-second time interval ($\delta t = 2$) between p_3 and p_4.

Compared to instantaneous movement derivatives, movement derivatives at the interval level are calculated taking into account both the spatial and temporal dimensions. A distance or time window moves along a given GPS trajectory to calculate interval movement derivatives including average velocity and maximum acceleration considering all GPS points within the range of the distance or time window. The calculated interval derivatives particularly reflect distinguishable spatial and temporal variations in the movement characteristics among different transport modes, taking into account traffic conditions. For example, assuming that the travel mode of the given GPS trajectory in Figure 2.3 is 'car' in smooth traffic flow, the average velocity of the GPS point p_7 within the 10-meter distance window is calculated based on the velocity values of p_6, p_7, and p_8 whereas the average velocity of the same GPS point p_7 within the 10 second time window is calculated based on the velocity values of p_5, p_6, p_7, p_8, and p_9. However, if the traffic was heavy around p_7, the number of GPS points considered for the 10-meter distance window would be more than the three points due to the many stationary GPS points near p_7 even though the size of the window is still the same as that in a smooth traffic flow. On the other hand, the 10-second time window is likely to shrink because of many stationary GPS points around p_7 in such heavy traffic, which can result in a drastic drop in average velocity. In this way, interval movement derivatives using distance and time windows can capture different focal characteristics of each transport mode in a complementary manner, in order to improve performance.

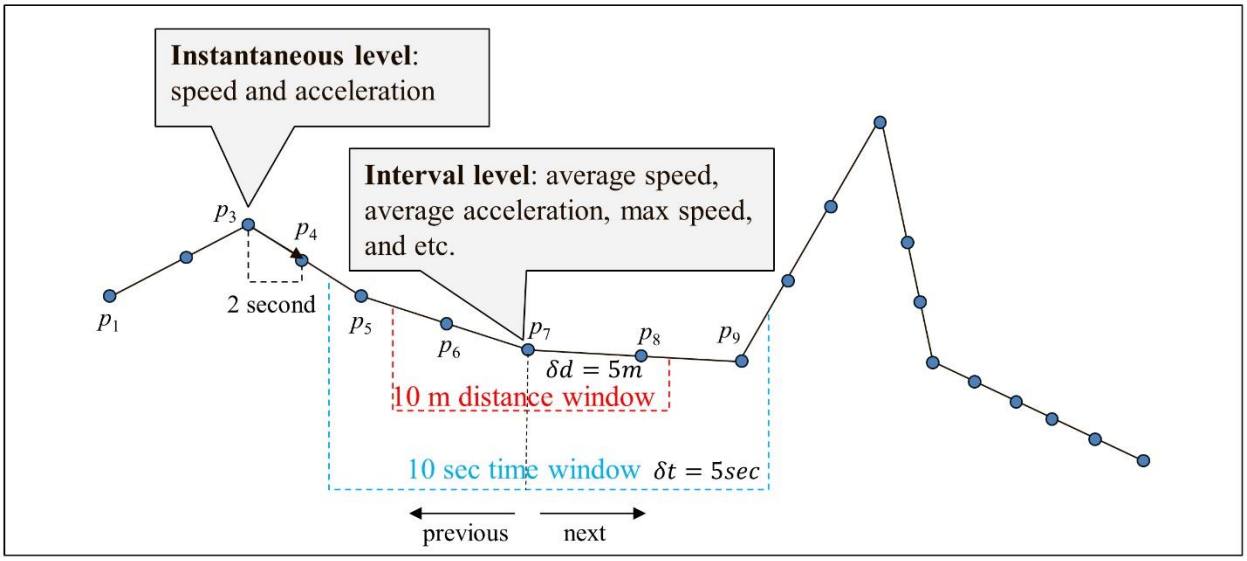


Figure 2.3. Movement derivatives calculated at instantaneous and interval levels

2.3.3 Indoor/outdoor classification and pre-processing of outdoor GPS points

Due to the limited role of indoor GPS data, outdoor GPS points need to be seperated from indoor GPS points. XGB (Chen & Guestrin, 2016) that follows the principles of gradient boosting (Friedman, 2001) is used to classify these two labels (indoor versus outdoor) based on three selected features derived from the GPS data collected from three subjects. Boosting is a forward stage-wise optimization approach that uses votes of each weak classifier, which is learned at every iteration, to generate a strong classifier. Gradient boosting uses tree models as weak classifiers and generates a strong classifier based on the notion of gradients in a way that a loss function is minimized. XGB especially takes regularization into account to control over-fitting for improved performance. Velocity, acceleration, HDOP, and the number of missing points are taken into account in feature extraction because GPS points may have a considerable amount of errors that manifest as position and velocity spikes and may not have been properly recorded due to signal loss. Only three important features — average velocity and HDOP at the 180-sec interval level, and the number of missing points at the 120-sec interval level — are

selected among all 31 features considering both instantaneous and interval levels of the GPS trajectory operators for the indoor/outdoor classification.

Preprocessing of GPS data involves determining valid indoor and outdoor GPS points. For the identified indoor GPS points, longitude and latitude values with high HDOP (e.g., HDOP > 8 in this study) are replaced with those of previous GPS points with low HDOP so that many arbitrarily dispersed GPS spikes can be displaced on a previously tracked place. Missing indoor GPS points due to signal loss are substituted with the positions of previous GPS points with the indoor label. Regarding the identified outdoor GPS points, GPS points with high HDOP are replaced with those estimated using Kalman filter to filter out noisy records and to produce accurate GPS points. The application of Kalman filter is described in Section 2.4.1.

2.3.4 Classification using GPS data

Since the Geolife GPS data are partitioned into many segments for each person and travel mode, two or more continuous trips with less than a 1-minute time gap need to be merged into one segment as an analytic unit for feature extraction. The merging process considers the transition between two different modes, like walking and riding a bus, for improving the identification of travel modes. For example, Person 1 and Person 2 traveled with different travel modes on 04/03/2008 and 08/27/2011, and different trips are divided into segments as shown in Table 2.3. In this case, because the first three segments of Person 1 from the first row to the third row happened continuously, they are concatenated into one segment. Further, since an important goal of the classification using the GPS data is to accurately separate traveling in a vehicle (the in-vehicle mode) and biking from other modes (like running) for better classification in the next phase using accelerometer data, consecutive segments of walking/running and other travel modes are merged into one concatenated segment for feature extraction.

Table 2.3. Examples of segments in Geolife GPS data

Row ID	Person ID	Start time	End time	Travel mode
1	**1**	**04/03/2008 11:32:24**	**04/03/2008 11:46:14**	**walk**
2	**1**	**04/03/2008 11:47:14**	**04/03/2008 11:55:07**	**taxi**
3	**1**	**04/03/2008 11:55:24**	**04/03/2008 12:01:49**	**taxi**
4	1	04/03/2008 16:00:00	04/04/2008 04:13:22	train
...	...	...	...	...
101	2	08/27/2011 06:13:01	08/27/2011 08:01:37	walk
102	2	08/27/2011 15:01:59	08/27/2011 15:31:43	walk

In addition, valid GPS segments are selected using the following criteria: 1) segments that consist of GPS points with an average recorded time interval of less than 10 seconds, 2) segments with a sum of recorded time of more than three minutes. GPS points recorded at longer time intervals cannot ensure the consistency of the features, so the segments with GPS points with such long time intervals are excluded. Further, segments with recorded time of less than three minutes are excluded due to the use of the interval-level operator with the largest time window of three minutes.

At the instantaneous and interval levels, a total of 73 features are generated for each GPS point as shown in Table 2.4. Two movement derivatives, velocity and acceleration, are calculated at the instantaneous level, and a set of five movement derivatives — average velocity, average acceleration, maximum velocity, maximum acceleration, and rate of change in velocity (Zheng et al., 2010) — is calculated for each time window (10-second, 20-second, and so on) and distance window (10-meter, 20-meter, and so on) at the interval level. Besides instantaneous and interval derivatives, the recorded hour of each GPS point is also extracted and used as a feature due to its important role in separating time-constrained activities. For instance, people are likely to engage in biking and running during daytime, and the use of motorized transport modes like bus and car is usually at its peak during the rush hours.

RF (Breiman, 2001) is used to identify vehicle-based movement (traveling in a vehicle and biking) and non-vehicle-based human movement (human locomotion) based on features extracted from the Geolife GPS dataset. Mean decrease Gini index is measured for each feature to examine important features. The most important movement derivatives are average velocity and maximum velocity at the interval level, whereas no movement derivatives at the instantaneous level show high importance (features with bold fonts in Table 2.4).

Table 2.4. All 73 features for classification using GPS data (bold-font features: top 20 important features). Movement derivative: variables derived from GPS trajectories regarding movements

GPS trajectory operator	**Movement derivative**	
Instantaneous	velocity	
	acceleration	
	recorded time (hour)	
Interval	*Time window*	
	10sec	average velocity, average acceleration, max velocity, max acceleration, change rate of velocity
	20sec	**average velocity**, average acceleration, max velocity, max acceleration, change rate of velocity
	30sec	**average velocity**, average acceleration, max velocity, max acceleration, change rate of velocity
	60sec	**average velocity**, average acceleration, **max velocity**, max acceleration, change rate of velocity
	90sec	**average velocity**, average acceleration, **max velocity**, max acceleration, change rate of velocity
	120sec	**average velocity**, average acceleration, **max velocity**, max acceleration, **change rate of velocity**
	180sec	**average velocity**, average acceleration, **max velocity**, **max acceleration**, **change rate of velocity**
	Distance window	
	10m	average velocity, average acceleration, **max velocity**, max acceleration, change rate of velocity
	20m	average velocity, average acceleration, max velocity, max acceleration, change rate of velocity
	30m	average velocity, average acceleration, max velocity, max acceleration, change rate of velocity
	40m	**average velocity**, average acceleration, max velocity, max acceleration, change rate of velocity

Table 2.4 (cont.)

GPS trajectory operator	Movement derivative	
	50m	average velocity, average acceleration, **max velocity**, max acceleration, change rate of velocity
	100m	**average velocity**, average acceleration, **max velocity**, max acceleration, change rate of velocity
	200m	**average velocity**, average acceleration, **max velocity**, max acceleration, change rate of velocity

2.3.5 Classification using accelerometer data

Running, walking, standing, and sedentary status are identified in the classification using accelerometer data based on the approach developed by Lee and Kwan (2018), which has a 99.03% predictive accuracy when RF is used. The time range of indoor GPS points delineates the boundary within which such specific travel modes are identified using accelerometer data. Identified human locomotion in the classification using GPS data also goes through the classification component using accelerometer data to generate specific travel modes.

2.4 RESULT

2.4.1 Performance of travel mode classification algorithm on real-world data

The travel mode classification algorithm was implemented using R statistical computing software (RCore, 2013). R supports machine learning model packages called ‘xgboost’ (Chen and He, 2015) and ‘randomForest’ (Liaw and Wiener, 2002) for XGB and RF used in the indoor/outdoor classification and classification using GPS data respectively. Since the time and distance windows at the interval level demand intensive computation in the indoor/outdoor classification and classification using GPS data processes, the ‘doParallel’ package, which deploys parallel computing through exploiting multi-cores, was used to improve computational

performance (Analytics & Weston, 2014). For the extraction of features from GPS data at the instantaneous and interval levels, it approximately spent 7 hours with four cores and parallel processing.

The performance of the two learning models generated in the indoor/outdoor classification and classification using GPS data processes were evaluated using 10-fold cross-validation. For the classification using GPS data, only 218,342 instances were randomly sampled and used for the cross-validation taking into account the original proportion of each label due to considerable computation time when all instances are input. The indoor/outdoor classification model using XGB showed 99.56% predictive accuracy in total with 500 iterations, and both indoor and outdoor GPS points were classified correctly with over 99% accuracy as shown in Table 2.5.

Table 2.5. Confusion matrix of indoor/outdoor classification using extreme gradient boosting. Confusion matrix: a summary table to describe the performance of a classification algorithm

		Actual (expected) class	
		Indoor	**Outdoor**
Predicted (observed) class	**Indoor**	35,140	106
	Outdoor	147	22,740
Accuracy (%)		99.58	99.54

For the classification using GPS data, RF achieved a 94.47 % predictive accuracy with 500 trees. The confusion matrix in Table 2.6 demonstrates that human locomotion and in-vehicle classes are identified with a high accuracy of over 90%. Biking has the lowest classification

accuracy (77.90%), and particularly, 1,833 and 1,061 instances with biking labels were incorrectly predicted as human locomotion and in-vehicle, respectively. Human locomotion shows the highest predictive accuracy (96.57%) among the three classes, which is, in turn, expected to deliver more accurately classified human locomotion instances to be further identified as one of the four travel modes in the classification using accelerometer data.

Table 2.6. Confusion matrix of classification component using GPS data. Confusion matrix: a summary table to describe the performance of a classification algorithm

		Actual (expected) class		
		Human locomotion	**In-vehicle**	**Biking**
Predicted (observed) class	**Human locomotion**	89,593	5,755	1,833
	In-vehicle	2,949	106,466	1,061
	Biking	235	250	10,200
Accuracy (%)		96.57	94.66	77.90

To validate the proposed algorithm on real-world data, GPS and accelerometer data collected from three subjects in free-living conditions were tested. These three subjects are all male, young adults and undergraduate or graduate students with no health issues. College students are appropriate subjects, who may take advantage of various urban opportunities in urban areas. Particularly, the three subjects often engage in moderate to vigorous PA in their daily lives by walking on campuses, performing exercises, or running and biking around their residential areas. They were given smartphones or used their own smartphones to record both GPS tracks and accelerometer data during a 7-day, 8-day, or 28-day period as shown in Table 2.7. Subject 1 especially participated in data collection for a duration of almost one month to

explore more activities with predicted results through visualization. All three subjects were asked to collect GPS and accelerometer data carrying the phone in their right pants' pockets. However, Subject 3 carried the provided smartphone in one of his jacket pockets by mistake, which resulted in poor predictive accuracy in the classification using accelerometer data, so the collected accelerometer data from Subject 3 were not used in this study. In addition, the accelerometer data from Subject 2 initially collected with an inexpensive phone (LG Realm) for a week were all erroneous. All three axes had mere variations of values, and the predicted results were mostly sitting status even when the subject walked. Hence, Subject 2 was asked to collect data again using Samsung Galaxy Alpha, which led to better classification results. GPS trajectories were recorded at a 1-second interval with HDOP. In addition, activity diaries that recorded subjects' travel in detail were also collected. Since the developed algorithm needs to be evaluated based on very detailed records of the subjects' activities and travel modes, each subject was asked to elaborate his activity diaries for each second through checking a visualized GPS trajectory. The elaborated three-day activity diaries, GPS trajectories and accelerometer data were then used as a test dataset for validating the algorithm.

Table 2.7. Description of GPS and accelerometer data collected from three subjects under free-living conditions

Subject ID	Device	Data	Recording time
1	LG G3, Samsung Galaxy Alpha	GPS and accelerometer	28 days
2	Samsung Galaxy Alpha	GPS and accelerometer	8 days
3	LG Realm	GPS	7 days

Since some GPS points needed to be manipulated due to their low accuracy and inherent positional errors, the Kalman filter was used to accurately estimate the latitude and longitude values of invalid GPS points. The Kalman filter is one of the widely used and best performing

smoothing methods for mitigating GPS random errors that influence the accuracy of derived measures from the GPS points (Jun et al., 2006; Grewal et al., 2011). The function 'dlmSmooth' in R for Kalman filter was utilized in this study to estimate missing or excluded GPS points and smooth GPS trajectories (Petris, 2009). With regard to the Kalman filter, parameter values suggested by a simple dynamic linear model of Petris (2009) were applied for variance of observation noise and variance of systematic noise at default. Higher variances of systematic noise were also tested to observe the effects of weak (variance of systematic noise = 1.0) or strong (variance of systematic noise = 0.01) filters, which make the test GPS trajectories less or more deviated from the original records (Figure 2.4). Kalman filtering with a variance of systematic noise of 0.1 had the best predictive accuracy and thus was applied to the test GPS trajectories for further evaluation of the travel mode classification algorithm.

The performance of the algorithm on real-world GPS trajectories achieved 96.20% accuracy in the identification of the six travel modes (Table 2.8). The travel mode identified with the lowest predictive accuracy was running (69.98%). 189 GPS points were wrongly classified as biking in the classification using GPS data. Walking and standing were the two most accurate types in the results (98.25% and 97.83%). Sensitivity, specificity, positive predictive value, and negative predictive value for the algorithm were also calculated (see Table 2.9).

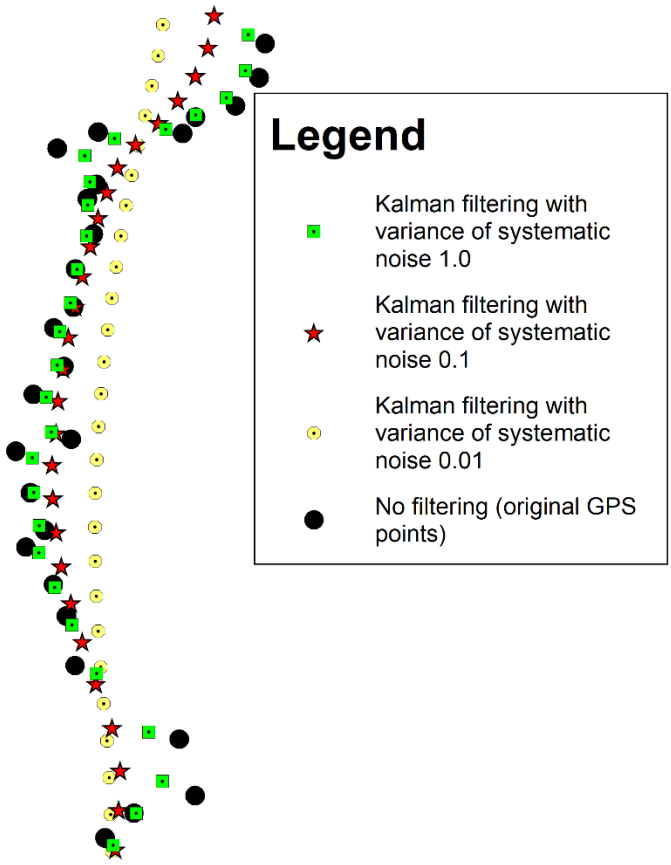


Figure 2.4. Testing Kalman filtering on GPS trajectories with three different variance values for systematic noise

Table 2.8. Confusion matrix of travel mode classification on free-living condition GPS and accelerometer dataset

		Actual (expected) class					
		Running	**Walking**	**Sitting**	**Standing**	**In-vehicle**	**Biking**
Predicted (observed) class	**Running**	464	19	0	0	0	0
	Walking	3	6725	1363	8	27	0
	Sitting	0	80	65889	0	101	2
	Standing	0	0	1019	2889	17	0
	In-vehicle	7	19	47	56	1662	175
	Biking	189	2	0	0	0	1779
Accuracy (%)		69.98	98.25	96.44	97.83	91.98	90.95

Table 2.9. Sensitivity, specificity, positive predictive value and negative predictive travel mode classification algorithm

Measures	**Running**	**Walking**	**Sitting**	**Standing**	**In-vehicle**	**Biking**
Sensitivity (%)	69.98	98.25	96.44	97.83	91.98	90.95
Specificity (%)	99.98	98.15	98.71	98.70	99.62	99.76
Positive predictive value (%)	96.07	82.76	99.72	73.61	84.54	90.31
Negative predictive value (%)	99.76	99.84	85.25	99.92	99.82	99.78

2.4.2 Predicted travel modes and its visualization with GPS trajectories

To quantitatively validate the travel mode classification algorithm, predicted results were combined with the collected GPS points based on the timestamp in both the accelerometer data and GPS trajectories (see Figure 2.5). GPS trajectories coupled with the predicted travel modes were visualized on a map using ArcGIS 10.4 to assess the predicted results of continuous trips and their relationships with ambient geographic contexts.

Car, bus, and biking mostly showed correct classification. Although there are waiting points with slow speeds around road intersections and bus stops, prediction of the car and bus showed promising results. However, compared to bus in the moderately urbanized area, some trajectories of traveling by bus in the urban area were incorrectly classified as biking. Biking also showed some wrong classification results as in-vehicle, which can be explained by its moderate predictive accuracy in Section 2.4.1. Walking, sedentary status (sitting), and standing were also accurately classified by RF through the classification using accelerometer data. Data on running were collected for a short time, like 10 minutes, from one subject, and some GPS points during running, however, were wrongly classified as biking. An above-ground subway trip, which the proposed algorithm did not take into account for the in-vehicle class, was detected as in-vehicle. A series of daily activities were also examined as shown in the right bottom of Figure 2.5. One subject went to a place by car for recreational purpose, and for this person, walking, sitting, and standing were represented on the recreational place indoors and outdoors.

Figure 2.5. Visualization of predicted six classified travel modes in free-living conditions. (a) car, (b) bus in moderate urban area, (c) bus in urban area, (d) biking, (e) running, (f) walking, (g): subway, (h) driving a car to a parking lot and performing recreational activities

2.5 DISCUSSION AND CONCLUSIONS

The hierarchical classification algorithm developed using publicly available GPS and accelerometer data in this study showed excellent performance in accurately predicting travel modes including in-vehicle status from real-world GPS and accelerometer data. Publicly available GPS and accelerometer data were shown to have potential for automatically and accurately classifying people's travel modes to conduct mobility research. The results also indicated that without any segmentation methods, in-vehicle status, biking, and other travel modes can be successfully classified with GPS trajectory operators. Based on our brief

exploration, it was also revealed that the relatively low accuracy of biking classification in Table 2.7 was due to some GPS points deviated from its original tracks presumably caused by positional errors and actual biking patterns with slow speed in business areas and possibly around destinations, even though I did not specifically analyze them in this study.

None of the instantaneous-level features were important in the classification using GPS data. This indicated that the interval-level movement descriptors calculated based on different sizes of time and distance windows outperform the instantaneous-level movement descriptors calculated from two consecutive GPS points, which have been widely used in many transport mode classification studies. In addition, the top 20 important features in Table 2.4 and 2.5 suggested that not only the time dimension but also the spatial dimension should be considered to extract features. Average and maximum velocity calculated through the time and distance windows played an important role in better predicting in-vehicle status, biking, and human locomotion.

Using heterogeneous sensor datasets to classify different travel modes with hierarchical processes is a novel approach and has considerable potential for application in a wide range of domains. For instance, Prelipccean et al. (2017) highlighted that interdisciplinary solutions regarding travel mode detection should not be limited to one research domain. They also emphasized that the current research trend of validating new algorithms and using datasets that cannot be shared widely hinders interdisciplinary studies. In this regard, this study provided a guidance for researchers in transportation and PA research to design a classification algorithm for travel modes in an extensible manner by training models using publicly available heterogeneous sensor datasets. In addition, I will make the algorithm available to interested researchers or practitioners upon request via a webpage. It is also imperative that predicted travel

modes with GPS trajectories potentially enable the further prediction of other activity-related or contextual information, including the activity itself and activity purposes, useful for addressing the UGCoP. For instance, GPS points of a subject around a bus stop, coupled with predicted 'standing' or 'sitting' can be interpreted as waiting time for a bus. Predicted labels 'walking' can also be further separated by its purpose through systematic assumption and logical reasoning if coupled GPS tracks are along certain contexts with specific purposes, like recreational facilities. Such inferred information by the predicted results, therefore, can help advance PA and transportation research. In addition, this study provides insights into how travel mode classification can be applied to various research domains involving human mobility analysis, such as air quality (Tainio et al., 2016), especially with regard to how the algorithm works on sensor datasets collected in free-living conditions and how the predicted results are represented.

In this study, GPS and accelerometer data collected from three subjects were classified using an algorithm based on the hierarchical classification processes. The classification results were visualized to enhance understanding of the subjects' daily activities and travel. The algorithm classified all activities with a series of plausible types of travel with high accuracy: riding in a car or bus in the moderately urbanized area, biking and sitting in a building, and standing around a bus stop. Especially, the interval level of the GPS trajectory operators in the classification using GPS data helped capture the characteristics of car and bus riding within different time and distance windows. It contributed to the correct classification of waiting points with slow speeds around road intersections and bus stops. Further, uncontrolled movement during a recreational activity was plausibly interpreted as a combination of different postures, like walking, standing, and sitting, which were characteristic of the activity. More important, the algorithm was able to detect indoor activities with high predictive accuracy.

The high performance of the developed algorithm, however, can be achieved only when the quality of collected sensor data is reliable. Low quality of built-in accelerometer sensors is likely to record erratic acceleration values, which may not be helpful for calculating features for accurate travel mode classification (Lee and Kwan 2017). The placement of smartphones can also greatly affect the result of classification using accelerometer data. Therefore, a smartphone screening procedure and instruction for participants on the placement of smartphones will be necessary to assure the performance of the algorithm. Additionally, none of the publicly accessible data used to train the models includes any personal information, like demographic information or socio-economic status, which indicates the specific population subgroups for whom the algorithm works best. Personal-level characteristics can account for internal variations among individuals in terms of not only age, gender, and race, but also daily PA patterns, health conditions, and transport modes, which in turn collectively constitutes population characteristics of the training data. However, since the Geolife GPS dataset, OpenStreetMap GPS dataset, and Wireless sensor data mining accelerometer data do not provide any individual or population characteristics, whether the algorithm can achieve the same performance for other population groups is uncertain.

The classification algorithm developed in this study has some limitations that need to be addressed in future research. First, additional processing should be applied in order to minimize incorrectly classified travel modes. For instance, bus riding in urban areas has inconsistent classification results whereas bus riding in moderately urbanized areas was perfectly classified. The incorrect classifications were largely the result of the low positional accuracy of the GPS points. In addition, a few wrongly predicted classes also appeared intermittently in the middle of

a trip. In this case, post-processing is needed to replace the erroneous travel modes with the correct ones.

In addition, the design of the classification algorithm needs to be improved for higher predictive accuracy. The hierarchical classification propagated an error downwards onto other classification processes. For example, in the travel mode prediction using the accelerometer data, some points of running were previously misclassified as biking in the classification using GPS data, and thus, the classification using accelerometer data could not intervene in the prediction and change the biking label to the running label even though the PA classification using accelerometer data correctly classfied it. Thus, future research should address how effectively the predicted results from the hierarchical classification processes can be exploited.

Finally, the motorized travel mode should be further classified to account for different motorized modes. This study combined train, bus, car, and taxi into a single motorized travel mode because the subdivision of the motorized transport does not have any significant meaning in terms of intensity and types of PA. However, for other research domains such as transport or social science, the differentiation of motorized travel modes, whether public or private, can be important (Biljecki et al., 2013; Feng & Timmermans, 2013; Shafique & Hato, 2016). Thus, the single motorized travel mode in this study needs to be subdivided and specified for this kind of research in the future.

2.6 REFERENCES

Almanza, E., Jerrett, M., Dunton, G., Seto, E., Pentz, M. A. (2012). A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. *Health & Place*, 18, 46–54.

Analytics, R., & Weston, S. (2014). doParallel: Foreach parallel adaptor for the parallel package. *R package version 1.0.8*, 1.

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J. L. (2012). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. *In International Workshop on Ambient Assisted Living*, 216–223.

Arif, M., Bilal, M., Kattan, A., Ahamed, S. I. (2014). Better physical activity classification using smartphone acceleration sensor. *Journal of Medical Systems*, 8, 95.

Biljecki, F., Ledoux, H. & Van Oosterom, P. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27, 385–407.

Bolbol, A., Cheng, T., Tsapakis, I. & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36, 526–537.

Boruff, B. J., Nathan, A. & Nijɲnstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood'in place-based health research. *International journal of health geographics*, 11, 22.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.

Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and*

Practice, 46, 1730–1740.

Browning, M. & Lee, K. (2017). Within what distance does "Greenness" best predict physical health? A systematic review of articles with GIS buffer analyses across the lifespan. *International Journal of Environmental Research and Public Health*, 14, 675.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, T. & He, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1–4

Cooper, A. R., Page, A. S., Wheeler, B. W., Griew, P., Davis, L., Hillsdon, M. & Jago, R. (2010). Mapping the walk to school using accelerometry combined with a global positioning system. *American Journal of Preventive Medicine*, 38, 178–183.

Council on Tall Buildings and Urban Habitat (2018). The skyscraper center. Retrieved from https://www.skyscrapercenter.com/city/chicago

Diez Roux, A. V., Mair, C., 2010. Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186, 125–145.

van Dijk, J. (2018). Identifying activity-travel points from GPS-data with multiple moving windows. *Computers, Environment and Urban Systems*, 70, 84–101.

Dodge, S., Bohrer, G., Weinzierl, R., Davidson, S. C., Kays, R., Douglas, D., Cruz, S., Han, J., Brandes, D., & Wikelski, M. (2013). The environmental-data automated track annotation (Env-DATA) system: linking animal tracks with environmental data. *Movement Ecology*, 1, 3.

Dodge, S., Weibel, R., & Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects.

Computers, Environment and Urban Systems, 33, 419–434.

Downs, J. A. & Horner, M. W. (2012). Probabilistic potential path trees for visualizing and analyzing vehicle tracking data. *Journal of Transport Geography*, 23, 72–80.

Dumais, S. & Chen, H. (2000). Hierarchical classification of Web content. *In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 256–263.

Eagle, N., Pentland, A. S. & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106, 15274–15278.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J. & Kerr, J. (2014). Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Public Health*, 2, 39–46.

Evenson, K. R., & Furberg, R. D. (2017). Moves app: a digital diary to track physical activity and location. *British Journal of Sports Medicine*, 51, 1169-1170.

Fang, S.-H., Liao, H.-H., Fei, Y.-X., Chen, K.-H., Huang, J.-W., Lu, Y.-D. & Tsao, Y. (2016). Transportation modes classification using sensors on smartphones. *Sensors*, 16, 1324.

Feng, T. & Timmermans, H. J. (2013). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies*, 37, 118–130.

Ferreira, N., Poco, J., Vo, H. T., Freire, J. & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19, 2149–2158.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of*

statistics, 29, 1189–1232.

Gong, H., Chen, C., Bialostozky, E. & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36, 131–139.

Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453, 779–782.

Gordon-Larsen, P., Nelson, M. C. & Beam, K. (2005). Associations among active transportation, physical activity, and weight status in young adults. *Obesity Research*, 13, 868–875.

Jankowska, M. M., Schipperijn, J. & Kerr, J. (2015). A framework for using GPS data in physical activity and sedentary behavior studies. *Exercise and Sport Sciences Reviews*, 43, 48.

Jansen, M., Ettema, D., Pierik, F. & Dijst, M. (2016). Sports facilities, shopping centers or homes: What locations are important for adults' physical activity? A cross-sectional study. *International Journal of Environmental Research and Public Health*, 13, 287.

Kwan, M.-P. (2004). GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler B*, 86, 267–280.

Kwan, M.-P., (2012a). How GIS can help address the uncertain geographic context problem in social science research. *Annals of GIS*, 18, 245–255.

Kwan, M.-P. (2012b). The uncertain geographic context problem. *Annals of the Association of American Geographers*, 102, 958–968.

Kwan, M.-P. (2013). Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers*, 103, 1078–1086.

Kwapisz, J. R., Weiss, G. M. & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter,* 12, 74–82.

Lachowycz, K., Jones, A. P., Page, A. S., Wheeler, B. W. & Cooper, A. R. (2012). What can global positioning systems tell us about the contribution of different types of urban greenspace to children's physical activity?. *Health & Place,* 18, 586–594.

Lam, M. S., Godbole, S., Chen, J., Oliver, M., Badland, H., Marshall, S. J., Kelly, P., Foster, C., Doherty, A. & Kerr, J. (2013). Measuring time spent outdoors using a wearable camera and GPS. *In Proceedings of the 4th International SenseCam & Pervasive Imaging Conference*, 1–7.

Laube, P., Dennis, T., Forer, P. & Walker, M. (2007). Movement beyond the snapshot–dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, 31, 481–501.

Lee, K. & Kwan, M.-P. (2017). Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted result. *Computers, Environment and Urban Systems*, 67, 124-131.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. *R News*, 2, 18–22.

Loukaitou-Sideris, A. (2006). Is it safe to walk? 1 neighborhood safety and security considerations and their effects on walking. *CPL Bibliography*, 20, 219–232.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K. & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing,* 23, 35–59.

Moiseeva, A., Jessurun, J. & Timmermans, H. (2010). Semiautomatic imputation of activity travel diaries: use of global positioning system traces, prompted recall, and context-sensitive learning algorithms. *Transportation Research Record: Journal of the*

Transportation Research Board, 2183, 60–68.

Outdoor Foundation (2016). Outdoor recreation participation topline report 2016. Retrieved from http://www.outdoorfoundation.org/pdf/ResearchParticipation2016Topline.pdf.

Papinski, D., Scott, D. M. & Doherty, S. T. (2009). Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12, 347–358.

Patrick, K., Kerr, J., Norman, G., Ryan, S., Sallis, J., Krueger, I., Griswold, W., Demchak, B., Dietrich, S., Raab, F. & others (2008). Geospatial measurement & analysis of physical activity: physical activity location measurement system (PALMS). *Epidemiology*, 19, S63.

Perchoux, C., Chaix, B., Cummins, S. & Kestens, Y. (2013). Conceptualization and measurement of environmental exposure in epidemiology: accounting for activity space related to daily mobility. *Health & Place*, 21, 86–93.

Petris, G. (2009). dlm: an R package for Bayesian analysis of dynamic linear models. University of Arkansas. Retrieved from ftp://ftp.math.ethz.ch/sfs/pub/Software/RCRAN/web/packages/dlm/vignettes/dlm.pdf

Prelipcean, A. C., Gidyfalvi, G. & Susilo, Y. O. (2017). Transportation mode detection–an in-depth review of applicability and reliability. *Transport Reviews*, 37, 442–464.

RCore, T. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online: http://www. R-project. org.

Rodriguez, D. A., Cho, G.-H., Evenson, K. R., Conway, T. L., Cohen, D., Ghosh-Dastidar, B., Pickrel, J. L., Veblen-Mortenson, S. & Lytle, L. A. (2012). Out and about: association of the built environment with physical activity behaviors of adolescent females. *Health &*

Place, 18, 55–62.

Sallis, J., Bauman, A. & Pratt, M. (1998). Environmental and policy interventions to promote physical activity. *American Journal of Preventive Medicine*, 15, 379–397.

Schrank, D., Eisele, B., Lomax, T., & Bak, J. (2015). 2015 urban mobility scorecard. Retrieved from https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015.pdf

Shafique, M. A. & Hato, E. (2016). Travel mode detection with varying smartphone data collection frequencies. *Sensors*, 16, 716.

Shoval, N. & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, 34, 141–159.

Shoval, N., Wahl, H.-W., Auslander, G., Isaacson, M., Oswald, F., Edry, T., Landau, R. & Heinik, J. (2011). Use of the global positioning system to measure the out-of-home mobility of older adults with differing cognitive functioning. *Ageing and Society*, 31, 849–869.

Tainio, M., de Nazelle, A. J., Gцtschi, T., Kahlmeier, S., Rojas-Rueda, D., Nieuwenhuijsen, M. J., de Sб, T. H., Kelly, P. & Woodcock, J. (2016). Can air pollution negate the health benefits of cycling and walking?. *Preventive Medicine*, 87, 233–236.

Tandon, P. S., Saelens, B. E., Zhou, C., Kerr, J. & Christakis, D. A. (2013). Indoor versus outdoor time in preschoolers at child care. *American Journal of Preventive Medicine*, 44, 85–88.

Tomlin, C. D. (1990). *Geographic information systems and cartographic modeling*. Englewood Cliffs: Prentice Hall.

Troped, P. J., Wilson, J. S., Matthews, C. E., Cromley, E. K. & Melly, S. J. (2010). The built environment and location-based physical activity. *American Journal of Preventive*

Medicine, 38, 429–438.

United States Census Bureau (2010). 2010 Census summary of Champaign County, Illinois. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml?src=bkmk

Voss, C., Sims-Gould, J., Ashe, M. C., McKay, H. A., Pugh, C. & Winters, M. (2016). Public transit use and physical activity in community-dwelling older adults: Combining GPS and accelerometry to assess transportation-related physical activity. *Journal of Transport & Health*, 3, 191–199.

Wang, J., Kwan, M.-P. & Chai, Y. (2018). An innovative context-based crystal-growth activity space method for environmental exposure assessment: A study using GIS and GPS trajectory data collected in Chicago. *International Journal of Environmental Research and Public Health*, 15, 703.

Weiss, G. M., Lockhart, J. W., Pulickal, T. T., McHugh, P. T., Ronan, I. H. & Timko, J. L. (2016). Actitracker: a smartphone-based activity recognition system for improving health and well-being. In *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 682–688.

Witayangkurn, A., Horanont, T., Ono, N., Sekimoto, Y. & Shibasaki, R. (2013). Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone. In *Proceedings of the international conference on computers in urban planning and urban management (CUPUM 2013)*, 1–19.

Xiao, Z., Wang, Y., Fu, K. & Wu, F. (2017). Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS International Journal of Geo-Information*, 6, 57.

Zhang, S., McCullagh, P., Nugent, C. & Zheng, H. (2010). Activity monitoring using a smart phone's accelerometer with hierarchical classification. In *Proceedings of the 2010 Sixth International Conference on Intelligent Environments (IE)*, 158–163.

Zheng, Y., Chen, Y., Li, Q., Xie, X. & Ma, W.-Y. (2010). Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web (TWEB)*, 4, 1.

Zheng, Y., Li, Q., Chen, Y., Xie, X. & Ma, W.-Y. (2008). Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on ubiquitous computing*, 312–321.

Zhou, X. (2014). Investigating the association between the built environment and active travel of young adults using location based technology. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Zhu, Q., Zhu, M., Li, M., Fu, M., Huang, Z., Gan, Q. & Zhou, Z. (2016). Identifying Transportation Modes from Raw GPS Data. In *Proceedings of the international conference of Young Computer Scientists, Engineers and Educators*, 395–409.

CHAPTER 3: BUFFER ANALYSIS AND ITS SIZE TO BEST PREDICT TRAVEL MODES FROM ENVIRONMENTAL CONTEXTS

3.1 INTRODUCTION

With the increasing interest in health behaviors and environmental contexts, many physical activity (PA) researchers have investigated the effects of physical and social environments on moderate to vigorous PA. For the investigation of the associations between PA and environmental factors, a large quantity of research used buffer analysis as one of GIS analysis methods, due to its simplicity of implementation and effectiveness in estimating the contextual influence within defined zones. The role of buffer zones with a pre-specified distance is to delineate areas within which individuals are potentially affected by environmental settings.

Due to the great contribution of residential neighborhoods to people's health, home addresses play an important role in defining buffer zones. For example, some studies investigated the effect of neighborhood green spaces around home locations on individuals' PA (Cohen et al., 2006; Coombes, Jones, & Hillsdon, 2010; Nagel et al., 2008; Schipperijn et al., 2013). In addition, for the last decade, since a growing body of research began to use GPS trajectories to detect specific places in which individuals perform moderate to vigorous PA for a certain amount of time, the use of buffer zones has been diversified to delineate areas for spatially immediate and temporally momentary influence of environmental settings along the GPS trajectories (Rodríguez et al., 2012; Burgoine et al., 2015).

In terms of the buffer size, many research groups sought a reasonable and reliable distance to capture true neighborhood effects. For instance, McGinn and colleagues (2007) showed that a 20-minute walk distance — roughly 1.6 km or 1 mile — was appropriate for

defining neighborhood areas from individual home locations in physical health research, whereas Berke et al. (2007) claimed that a slightly smaller size, 1 km or 0.6 miles, might better capture the characteristics of residential neighborhoods. Browning and Lee (2017) also found that greenness calculated within circular buffers between 500 m (0.31 miles) and 1999 m (1.24 miles) in size best predicted physical health.

However, a few studies have been conducted to address the methodological issue that might be raised by the buffer size with regard to the estimation of dynamic exposure along individuals' daily trips on GPS trajectories when compared to static locations, such as home addresses. For example, Rodríguez et al. (2012) created 50 m circular buffers around each GPS point to avoid the potential dependence on estimated effects of built environment between two consecutive GPS points, although the associations between environmental characteristics and moderate to vigorous PA that may vary depending on different buffer sizes were not closely examined. Due to a lack of consensus on the buffer distance in previous studies, findings were inconsistent; thus, the varying associations need to be investigated, taking into account buffers with different distances, to provide insights into understanding the implication of the parameter on research findings. In addition, it is important to suggest a reliable distance reflecting less variability in the effects and significance of the associations.

Thus, this study addresses a methodological issue on how the associations between environmental contexts and active travel modes (ATMs) vary depending on buffer sizes. Sensitivity analyses are conducted to investigate the varying associations between 7 environmental factors (e.g., green space, crime, traffic crashes) and walking and biking, taking into consideration 11 different buffer sizes ranging from 20 m to 200 m. Many studies especially reported mixed or non-significant findings regarding safety-related factors. This study includes

crime and traffic crashes as public safety factors (Hoehner et al., 2005; McGinn et al., 2007; Nagel et al., 2008; Troped et al., 2010; Boruff, Nathan, & Nijënstein, 2012). In addition, this study uses multinomial logistic regression to statistically examine the associations and variations of coefficient estimates, and the statistical significance of each predictor according to the 11 different buffer sizes are explored to find a reliable distance.

This study contributes to accurate estimation of direct exposures to environmental contexts. Buffer captures relatively more direct influence of environmental settings along GPS trajectories when compared to other GPS-based delineation methods, including activity space (Zenk et al., 2011; Hirsch et al., 2014; Perchoux et al., 2014; Lee et al., 2016; Rundle et al., 2016), kernel density (Thierry et al., 2013; Jankowska et al., 2017), and daily potential path area (Kwan, 1999). Further, buffers created over each GPS point take into account temporal weights, in terms of accumulative exposures to environments, when individuals stay for a longer time around specific places, whereas other delineation methods often neglect them (c.f., Wang & Kwan, 2018). With these buffer characteristics in mind, this study seeks to find a reliable buffer distance for mobility research in the fields of health and transportation to enable accurate estimation of direct exposure to the environment.

This study is structured as follows: past studies on the estimation of environmental exposure using buffers are reviewed in Section 3.2. GPS data, GIS data, and sensitive analyses depending on different buffer sizes are explained in Section 3.3. Section 3.4 shows the results from the statistical analyses on varying associations between ATMs and multiple environmental factors, and Section 3.5 discusses and concludes the research findings.

3.2 ESTIMATION OF ENVIRONMENTAL EXPOSURE USING BUFFER ANALYSIS

GIS analysis methods have been used to quantitatively estimate the impact of environments on PA. Especially, a large body of research used the buffer analysis method to delineate neighborhoods around home addresses and find empirical evidence regarding the associations between PA and neighborhood characteristics (Hillsdon et al., 2006; Berke et al., 2007; McGinn et al., 2007; Maas et al., 2008; Nagel et al., 2008). Buffer is one type of spatial analyses that can objectively assess the effects of various factors by defining zones around geometric primitives, like points, lines, or polygons, which represent geographic elements. Circular (radial) buffer is widely used to define neighborhood areas in an isotropic manner with the same distance in all directions, whereas network buffer is to define anisotropic neighborhood areas taking into account subjects' reachable distances along roads. A circular shape has been widely used to delineate neighborhoods based on a specific distance from a set of entities. McGinn and colleagues (2007) created 1.6 km (1 mile) circular buffers around every participant's residence to examine the associations between built environments and PA of adults for leisure and transportation purposes. The 1.6 km distance accounts for a 20-minute walking distance to delineate residential neighborhoods from a participant's home.

Because the neighborhood areas can be defined based on different characteristics of participants or environments, various buffer sizes were considered in some studies (Berke et al., 2007; Maas et al, 2008; Nagel et al., 2008; Mitchell et al., 2016; Chambers et al., 2017). Berke et al. (2007) especially used smaller sizes of buffers (e.g., 100, 500, and 1 km), which may better explain the potential walkable areas of older adults around their homes. As one of built environments, green space may have different levels of effects on PA depending on the proximity between green space and home. Maas and colleagues (2008) used 1 km and 3 km

radial buffers to measure the percentage of green space around participants' homes, while Cerin and colleagues (2017) applied 500 m and 1 km network buffers to define common reachable neighborhoods for adults over 12 countries. Browning & Lee (2017) conducted a systematic literature review and found that when the buffers were created around home addresses, the large number of studies showed significant associations between greenness and better physical health, including PA, as the buffer size increases up until 1999 m. Further, Nagel and colleagues (2008) concluded that the significance of the associations between some built environments and PA could vary depending on the sizes of buffers, which represent different ranges of neighborhood areas.

The increased adoption of GPS receivers in research has led to the identification of locations where PA occurs. In some studies, GPS points falling within 400 m to 1600 m radial or network buffers from home locations were considered (Almanza et al., 2012; Boruff et al., 2012; Dunton et al., 2014; James et al., 2014). Those studies combined GPS points with objectively measured PA to understand the effect of residential settings on PA. Boruff and colleagues (2012) further investigated different buffer types and their influence on research findings. Different sizes of buffers were also considered in recent studies to define more accurate residential neighborhoods for specific PA types (e.g., walking or bicycling) (Hirsch et al., 2016; Prins et al., 2014). In other words, in terms of the issue on the size, PA types became another factor to be considered beyond the general definition of residential neighborhoods in public health.

However, only a few studies have used buffer taking into account whole or partial individual trips traced by GPS trajectories for assessing more dynamic influence of environmental settings along the GPS trajectories. For instance, Rodríguez et al. (2012) justified the use of 50 m buffers around each GPS point to estimate daily exposures of adolescent females

to built environment characteristics. The purpose of using the 50 m distance was to avoid the potential dependence in the estimated effects of built environment between two continuous GPS points. Further, Burgoine and colleagues (2015) applied a hybrid method by using 100 m circular buffers for estimating environmental exposures during trips from and to school of children and 800 m network buffers for residential and school neighborhoods. Regarding the trips from and to school of children, Harrison and colleagues (2014) tested actual routes from GPS points and predicted routes calculated by the shortest path algorithm to compare the food and PA environments in the 100 m buffers along the two different kinds of routes. Yin and colleagues (2013) highlighted that although moderate to vigorous PA of youth usually occured within 0.25 or 0.3 mile radius around their residences considering their GPS trajectories, the patterns of the space-time paths was not uniformly distributed on the radial area.

Therefore, an in-depth investigation into the buffer size for PA research involving GPS trajectories is conducted in this study. This study will give ideas of whether the implication of buffer size exists and if so, how the distance affects the associations between PA and multiple physical and social environmental factors that previous studies haven't explored yet.

3.3 METHOD

3.3.1 GPS data description

GPS trajectory data with survey results and daily activity diaries obtained from the Chicago Regional Household Travel Inventory (CRHTI) project are used in this study. Chicago is the third largest metropolitan area in the U.S. where people are exposed to various urban opportunities and built environments. The study was approved by the University of Illinois Institutional Review Board. In the CRHTI survey, GPS trajectory data were recorded from

members of 147 households for 7 days between September 2007 and December 2007. Daily activity diaries were reported during the first day of the entire 7 days. Among these participants, only 178 persons from 73 households had both complete personal and household information in addition to complete GPS data and activity diaries, and 168 adults more than 18 years old are used in this study. The GPS trajectory data were recorded at a 5-second interval when a participant was moving at a speed of at least one mile per hour (mph), which is the speed of slow walking.

3.3.2 Travel mode classification

To obtain the travel modes of trips, the travel mode classification algorithm demonstrated in Chapter 2 was applied. Among the three main processes that constitute the travel mode classification algorithm, the "Classification using GPS data" process as part of the entire algorithm is modified, optimized, and implemented to classify walking, in-vehicle status, biking, and running for this study since GPS data were the only sensor data collected through the CRHTI project. The optimization process includes the test of larger ranges of distance and time windows than the already developed version to see if there is a big difference in predictive accuracy between small and large sizes of windows.

The overall predictive accuracy of the optimized classification was 97.00% (walking: 97.58%, in-vehicle: 97.24%, biking: 89.33%, running: 99.01% in Figure 3.1) with 10 to 300 sec time and 10 to 500 m distance windows when the Geolife GPS data described in Section 2.3.1 were used as training and test datasets. When 10 to 180 sec time and 10 to 200 m distance windows, which were the minimum range of window sizes in the test, were used, it showed the lowest predictive accuracy (95.04%). The optimized classification was also tested on the real-world GPS trajectories collected from three subjects described in Section 2.4.1. It improved the

accuracy of walking (99.38%) and in-vehicle status (95.81%), whereas biking (90.47%) showed a bit lower accuracy than the original version of algorithm. Running observations in the real-world GPS trajectories were replaced in this study with almost 2-hour records from TrackProfiler because the quantity of observations was too small to measure the performance of the algorithm in predicting running. With the replaced observations, the optimized classification showed 90.95% of accuracy in identifying running. Predicted results of the CRHTI GPS data and their clustering patterns are represented using kernel density estimation in Figure 3.2.

3.3.3 Statistical analyses

Circular buffers are created around each GPS point with the predicted travel modes (Figure 3.3). As for the impact of environments on PA, the in-vehicle status contrasts with active travel modes, like biking (Winters et al., 2010). Therefore, the in-vehicle status can serve as a reference category when the associations between each travel mode and environmental contexts are examined.

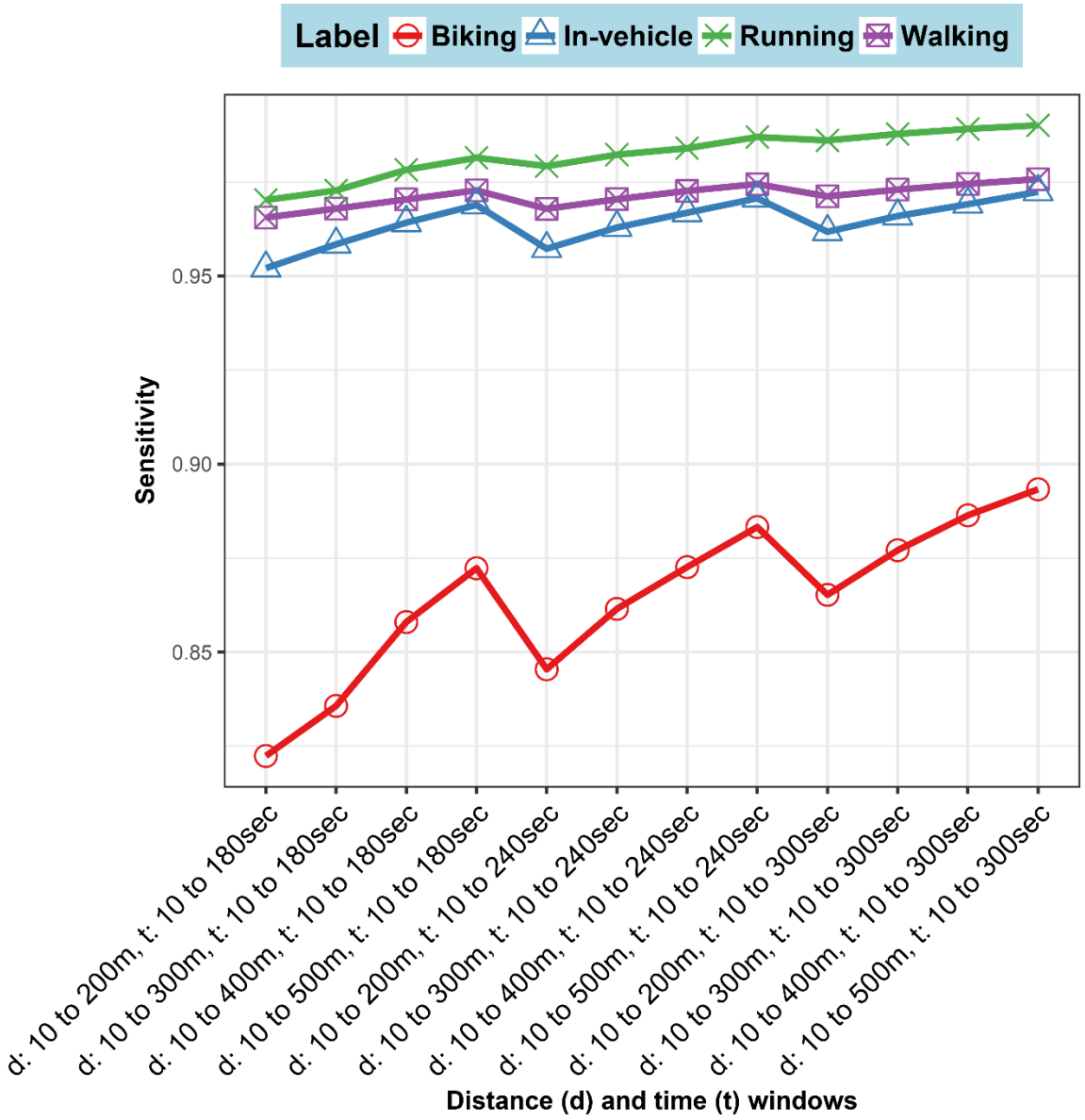
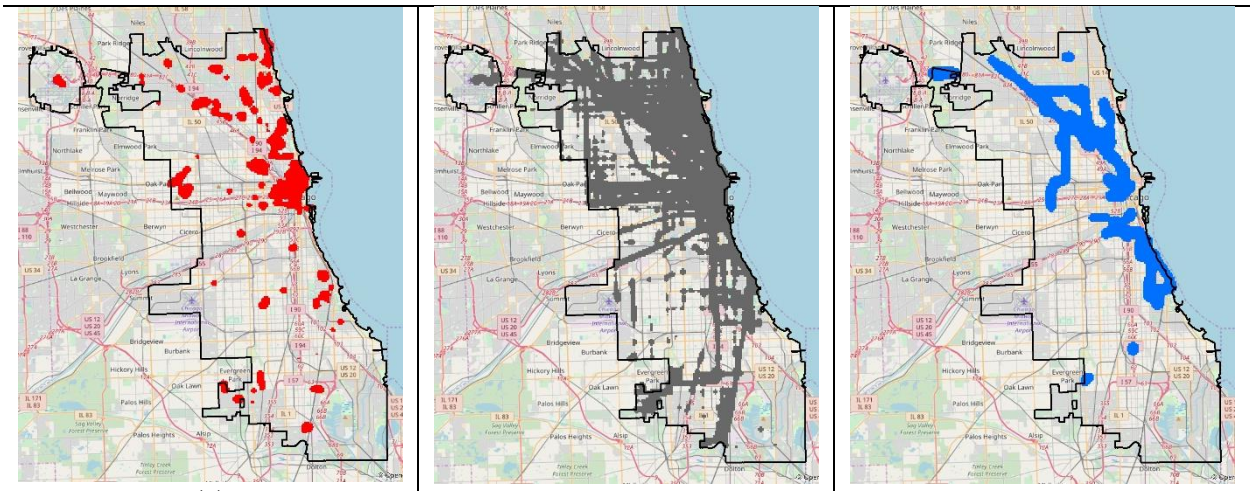


Figure 3.1. Performance test of optimized travel mode classification using GPS data with different combinations of features depending on the sizes of distance and time windows



(a) (b) (c)

Figure 3.2. Predicted travel modes on Chicago Regional Household Travel Inventory GPS data and their clustering visualization using kernel density estimation. (a) walking, (b) in-vehicle, (c) biking

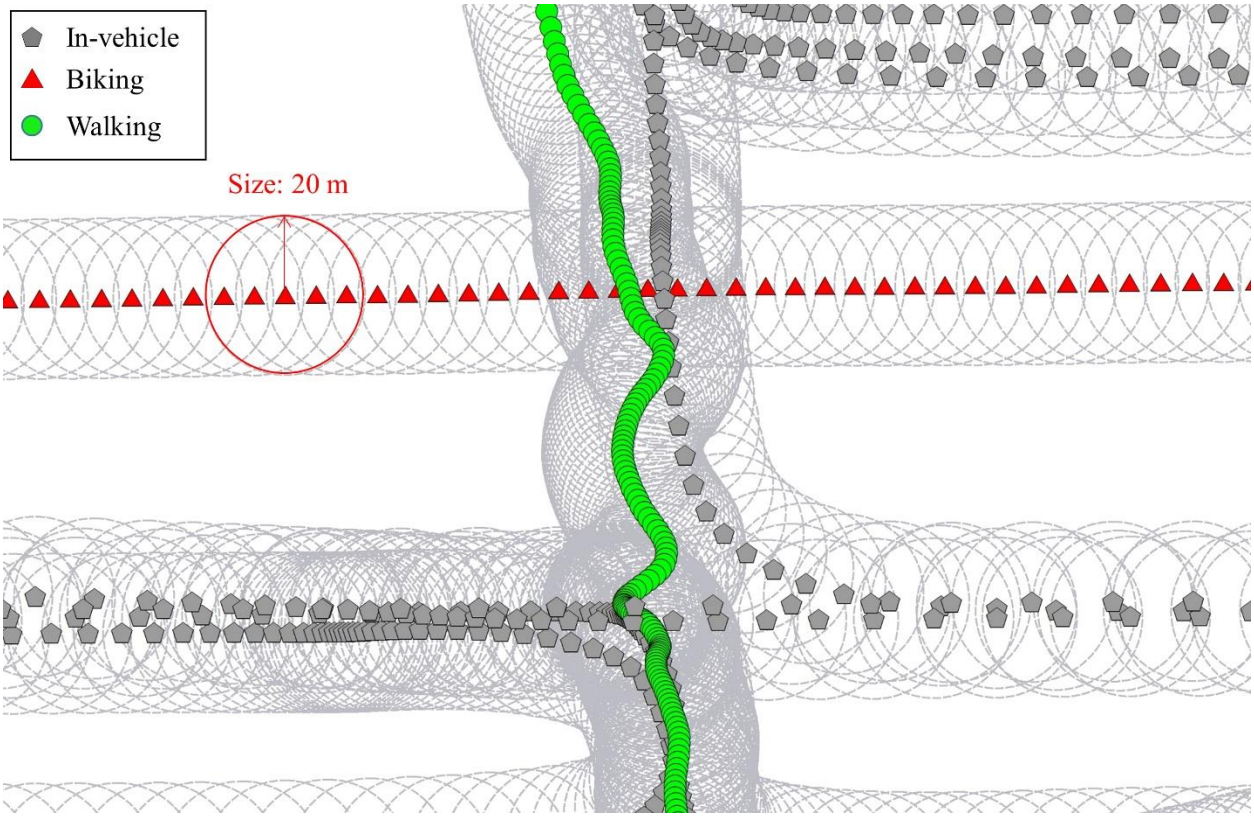


Figure 3.3. 20 m circular buffers created around each GPS point recorded at 1-second interval with predicted travel modes

The distance as one of its key properties in buffer analysis and its influence on the research findings is investigated. For the statistical model in this study, multinomial logistic regression is used to understand the associations between travel modes and environmental contexts. As of the response variable, predicted inactive and ATMs are compared (e.g., in-vehicle vs. walking, in-vehicle vs. biking). Logistic regression analyses provide odds ratios (ORs) for each predictor to know the probability that a specific outcome (e.g., active travel) can happen against another (e.g., inactive travel). Each observation considered in the models is a GPS point, and 7 different environmental variables are measured based on buffer areas specified by different distances and assigned to each observation. As shown in Table 3.1, those 7 predictors are the measures calculated from physical, social, and safety-related environmental contexts. The crime predictor includes 11 types of violent crimes, and the incidence of the crimes is estimated (Bureau of Justice Statistics, 2018; National Institute of Justice, 2018). Trees and parks and open spaces are included for greenness and land use respectively, and traffic crashes involving pedestrians and pedal cyclists related to public safety is taken into account in this study as well. Transit availability index is a measure that "takes into account transit service frequency, pedestrian friendliness, network distance to transit stops, and number of subzone connections." (CMAP Data Hub, 2017) and also incorporated. Aside from these predictors, age, race, household income, and weekday/weekend are confounded as individual characteristics.

Three models are iteratively generated on 11 different sizes — 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, and 200 m — of buffers, and therefore, a total of 33 models are created. Among the three models, a first model has the number of cases of violent crimes, percentage of tree areas, park and open space density, transit availability index, and traffic crashes adjusted for age, race, household income, and weekday/weekend. A second model adds neighborhood African

American as a predictor on top of the first model, and a third model additionally includes neighborhood median household income to understand how both individual and neighborhood characteristics — African American and household income in this study — can affect the likeliness of individuals towards walking or biking against traveling in vehicles.

R statistical software is used for the computation of variables on a large number of GPS points with different buffer sizes and statistical analyses. Especially, Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream with Intel Xeon E-2680v3 CPUs (24 cores) aids accelerating the variable calculation (Towns et al., 2014; Stewart et al., 2015).

Table 3.1. Description of 7 predictors measured for environmental contexts

Predictors	Data source	Measure	Measuring unit	Resolution/ unit	Time	Comments
Crime	Chicago Data Portal	Count	Number of crimes / km^2	Point	2007	Violent crimes
Greenness	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	Tree
Land use	Chicago Data Portal	Percentage	Area (m^2)	Polygon	2010	Park and open space
Neighborhood median household income	American Community Survey of United States Census Bureau	Average	Dollars ($)	Census tract (polygon)	2010	
Neighborhood population	American Community Survey of United States Census Bureau	Average	Number of people / km^2	Census tract (polygon)	2010	African American
Transit availability index	Chicago Metropolitan Agency for Planning Data Hub	Average	Access to Transit Index	Polygon	2010	
Traffic crash	Illinois Department of Transportation	Count	Number of crashes / km^2	Point	2007	Pedestrian and pedal cyclists

3.4 RESULT

3.4.1 Descriptive statistics

Table 3.2 shows the descriptive statistics of the 168 adults with their personal characteristics and predicted average daily travel time using the optimized travel mode classification algorithm. The percentage of females is slightly higher than males. Most of the participants are whites and middle-aged adults. According to the population statistics of Chicago (United States Census Bureau, 2010), the actual percentage of whites is 45% in the study area compared to 81.5 % in our sample, and African Americans account for 33%, which is higher than the percentage observed in our sample (10.7 %). High-income people comprise the dominant group in the sample, and vehicles including private cars and public transit are the most-used modes for daily travels, which accounts for one hour per day on average. Since running was rarely performed in the daily lives of the 168 participants, it is excluded for this study, and the other three travel modes (i.e., walking, traveling in a car, and biking) are considered.

A part of total GPS points is sampled at 10-second level for this study for efficient computation and analyses (n = 156,627). The environmental characteristics within 50 m buffers, which is widely used in the past studies, around the 156,627 GPS points are described in Table 3.3 to give a sense of how the samples are dynamically exposed to different environmental contexts in their daily lives. Regarding the predicted travel modes, 40,999 GPS points were identified as walking, whereas only 2,545 points were classified as biking. GPS points with predicted walking had higher tree density, higher transit availability and battery incidence, and more traffic crashes involving pedestrians and pedal cyclists at average than biking and in-vehicle. On the other hand, observations with the predicted biking had more park and open space areas and higher neighborhood median household income on average. The correlations among

the 7 predictors were analyzed to evaluate multicollinearity, and it was found that there was no high correlation in any pairs of the predictors (not present).

Table 3.2. Descriptive statistics of the 168 adult participants in the Chicago Regional Household Travel Inventory project

n = 168 persons	**Percentage (%)**
Female	53
Race	
White	81.5
African American	10.7
American Indian or Alaska Native	1.2
Asian	1.8
Hispanic	3.0
Other	1.8
Household income (73 households)	
< $20,000	8.7
$20,000 - $34,999	5.4
$35,000 - $49,999	3.3
$50,000 - $59,999	9.8
$60,000 to $74,999	10.9
$75,000 to $99,999	21.7
$100,000+	39.1
NA	1.1
	Mean ± standard deviation
Age	43.2±11.4
Predicted average daily travel time (hours)	
Walking	0.3±0.2
Running	0.0003±0.0
Biking	0.02±0.5
In-vehicle	1.1±0.7
Number of recorded days	5.3±1.5

Table 3.3. Descriptive statistics of 7 predictors measured around all GPS points using 50 m circular buffers

Predictors	Walking (n = 40,999)				Biking (n = 2,545)				In-vehicle (n = 113,083)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Physical environment												
Greenness - Tree (%)	18.00	19.14	0	100.00	9.00	10.49	0	52.60	11.80	13.15	0	100.00
Land use - Park and open space (%)	5.50	19.12	0	100.00	8.70	25.76	0	100.00	8.00	23.73	0	100.00
Transit availability index (1–5)	4.67	0.46	3.41	5.00	4.31	0.45	4.00	5.00	4.47	0.51	1.00	5.00
Social environment and safety												
Crime (count/km^2)	41.00	74.81	0	844.00	19.90	31.94	0	214.90	28.00	57.17	0	1146.00
Neighborhood median household income ($)	59,484	22,832	13,177	127,736	69,253	19,753	27,866	127,460	55,161	22,122	10,217	150,281
Neighborhood African American (population/km^2)	1,742	2,354	4	10,325	269	335	4	2,630	1,207	1,767	0	10,325
Traffic crash (count/km^2)	3.80	8.65	0	71.60	3.10	7.33	0	47.70	3.20	7.70	0	79.60

3.4.2 Sensitivity analyses of 11 sizes of buffers

The varying associations depending on 11 different sizes of buffers in Model 1, 2, and 3 are shown in Figure 3.4. In the three models, it was found that different buffer sizes affected the associations between ATMs and environmental factors in terms of significance levels of variables and ORs. When trees, parks and open spaces, transit availability, battery, and pedestrian traffic crashes were included as predictors in the Model 1, the associations were mostly significant in walking across 20 to 200 m buffers, whereas only trees and traffic crashes involving pedestrians and pedal cyclists had significant associations in biking consistently along different buffer sizes. When the neighborhood African American population was added as a predictor in the Model 2, the transit availability became more significant in all the buffer sizes in biking vs. in-vehicle; however, the parks and open spaces and crime still remained non-significant in most of buffer sizes on biking. The added neighborhood African American population was also not significantly associated with biking as well. On the contrary, the results of the Model 3 with the additional neighborhood median household income indicated that the implications of buffer sizes were eventually alleviated in tree, transit availability, and neighborhood household income regarding their significant levels in walking and biking. The rest of the predictors, however, did not show consistent significance levels along the buffer sizes. Particularly, as to the parks and open spaces, crime, and neighborhood African American population, only relatively large buffer sizes — 150 and 200 m — turned to be significant in biking. Further, the traffic crashes were not found to be significant in biking vs. in-vehicle until the buffer size reached to 40 m.

In all the three models, the associations of walking vs. in-vehicle with all predictors mostly had high significance levels ($p < 0.001$), and the ORs varied as the buffer size altered,

while the graphs of biking vs. in-vehicle mostly showed stable trends along the buffer sizes, except for crime and traffic crashes. In the two safety-related factors, the ORs of walking and biking compared to in-vehicle status especially showed common characteristics. They both began with similar ORs in low buffer sizes around 20 and 30 m, diverged more and more as the size became larger, and crossed at a certain size (crime) or were widened further (traffic crashes).

Since the Model 3 had more significant variables along different buffer sizes, and 200 m was the utmost distance showing the largest number of higher significance levels in all predictors in the Model 3 as shown in Figure 3.5, it was selected as the most appropriate buffer distance to examine the associations between ATMs and environmental factors in this study. In Figure 3.5, the histogram indicated that as distances got closer to the 200 m, the count of significant variables became higher in total.

Figure 3.4. Varying odds ratios and standard errors resulted from Model 1, 2, and 3 across 11 different sizes of buffers. Model 1: the number of cases of violent crimes, percentage of tree areas, park and open space density, transit availability index, and traffic crashes adjusted for age, race, household income, and weekday/weekend. Model 2: Model 1 + neighborhood African American as an additional predictor. Model 3: Model 2 + neighborhood median household income as an additional predictor.

Model 1

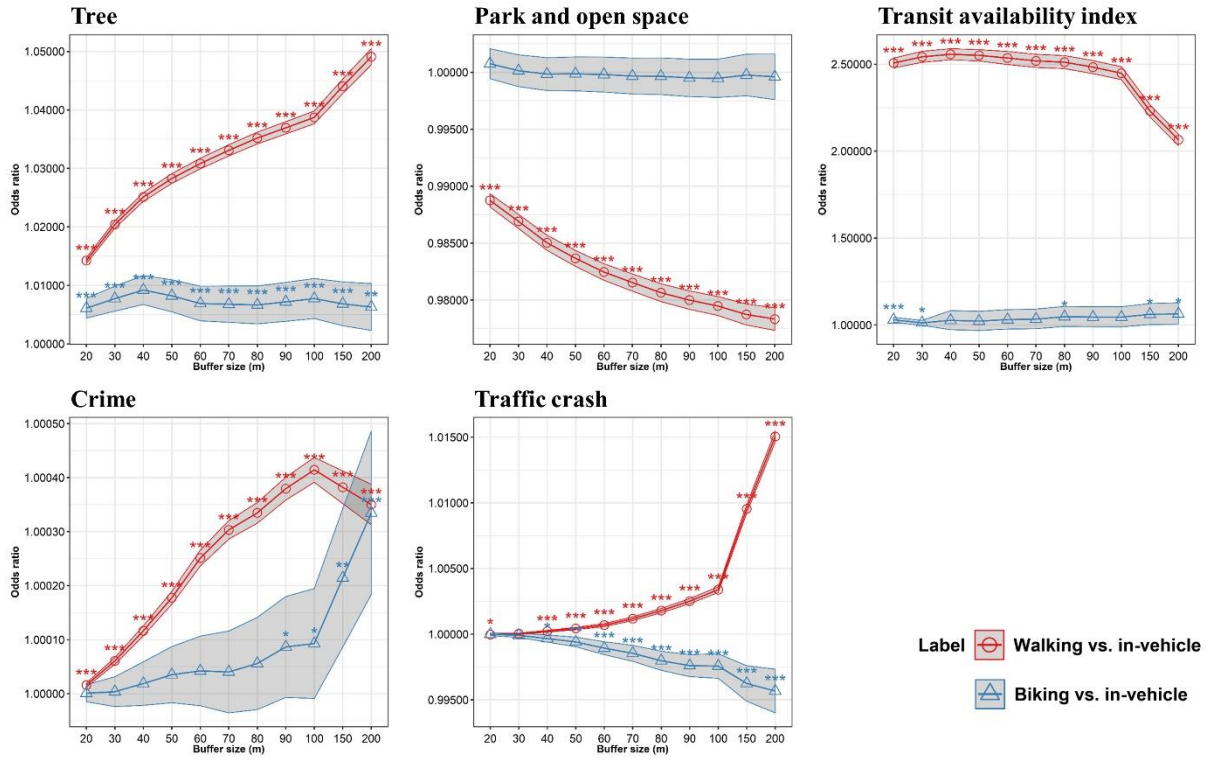


Figure 3.4 (cont.)

Model 2

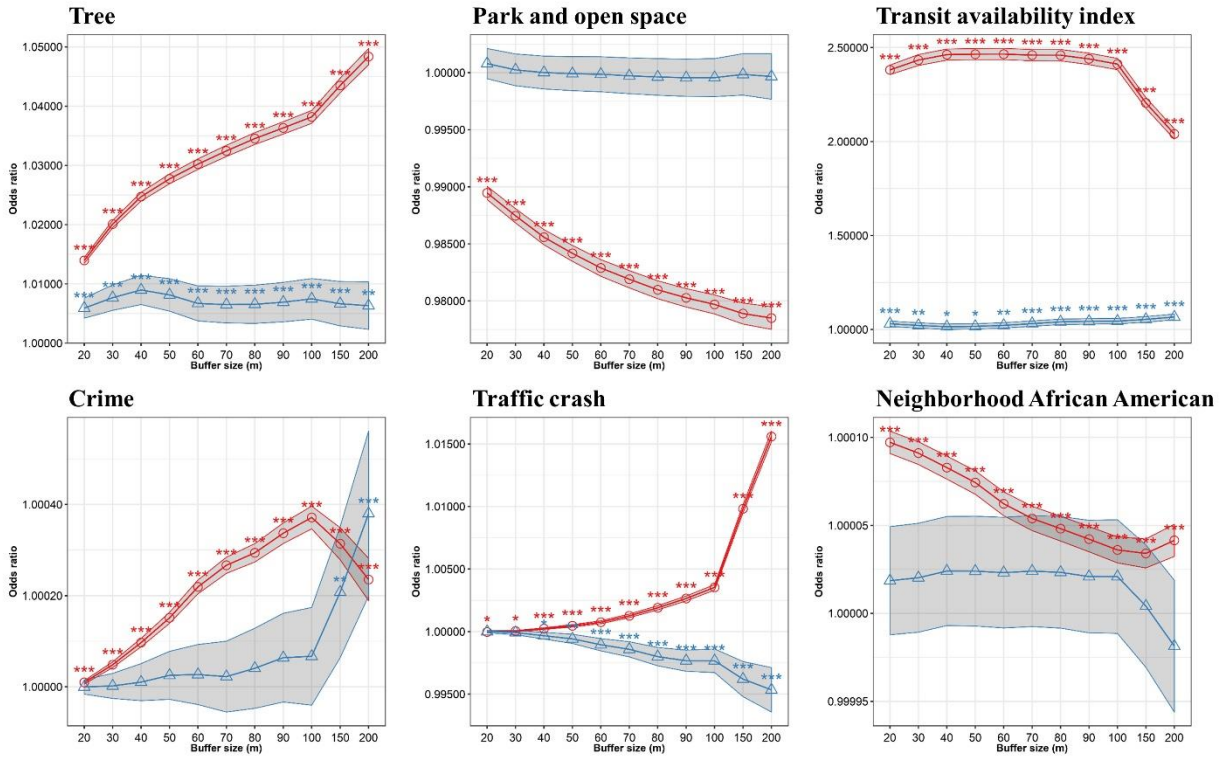
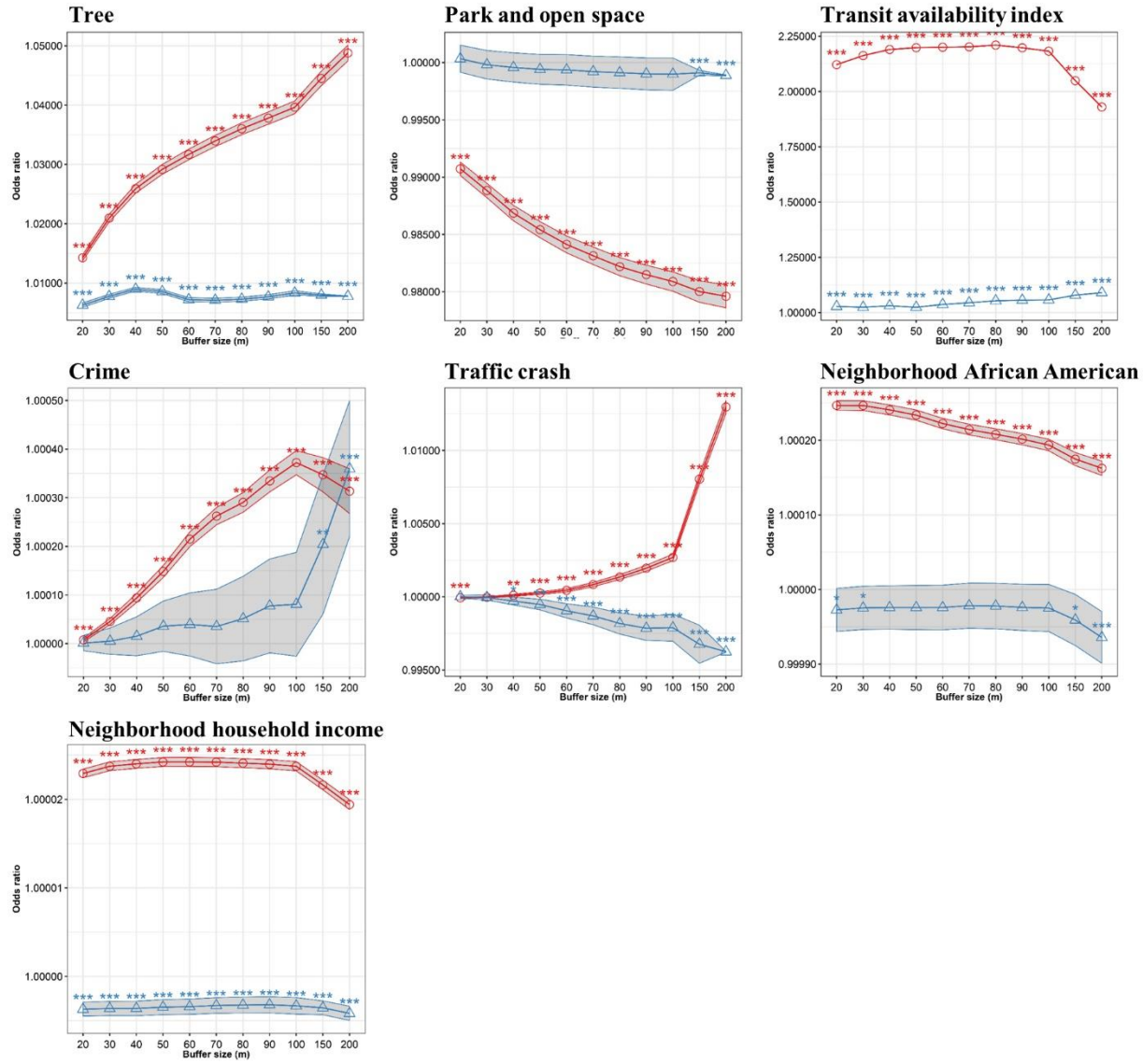


Figure 3.4 (cont.)

Model 3



***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.1$

Gray regions: standard errors of coefficients

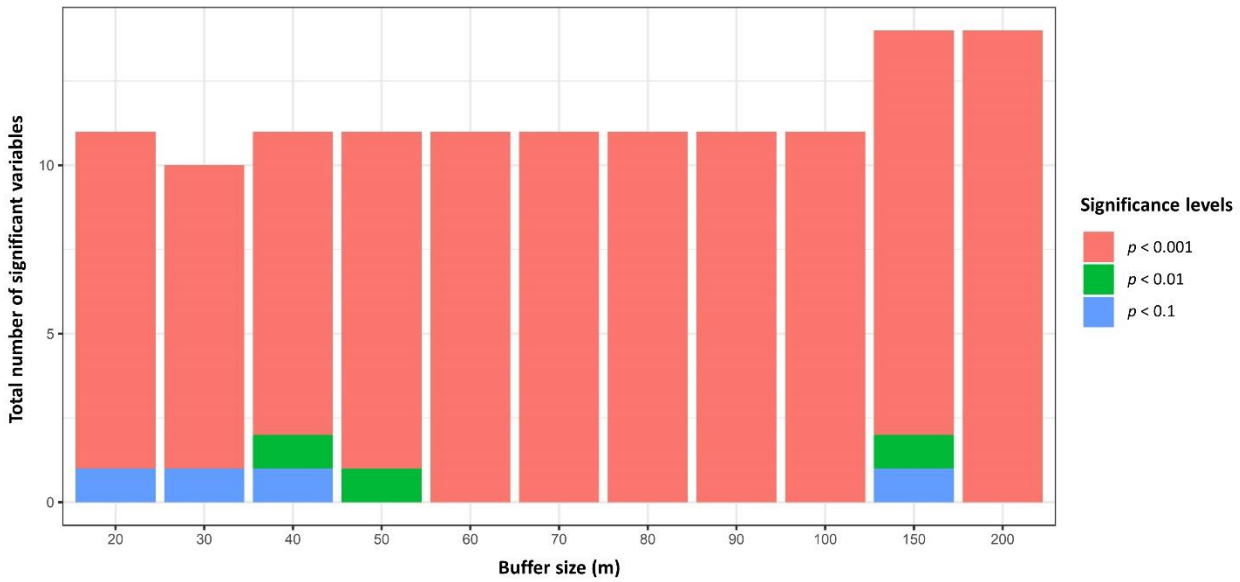


Figure 3.5. Total number of significant variables in Model 3 according to different buffer sizes

The associations of the 7 predictors measured within 200 m buffers of each GPS point and travel modes are shown in Table 3.4. All the model fit measures, including the three kinds of pseudo R-squared, indicated that the Model 3 with the added neighborhood household income better explained the variability of the predictors than other two models. The higher percentage of tree areas (OR: 1.05), transit availability (OR: 2.06), incidence of crime (OR: 1.00), and traffic crashes (OR: 1.01) were significantly associated with higher odds of walking, whereas the traffic crashes (OR: 0.99) were significantly associated with lower odds of biking compared to in-vehicle status in Model 1. For walking, the density of parks and open spaces (OR: 0.97) had a significant association with lower odds of walking against in-vehicle status in the Model 1. The results of Model 2 were similar to the Model 1, and the model fit measures of the Model 2 were not improved much after the neighborhood African American population was added.

Compared to the Model 1 and 2, all the variables had significant associations with walking and biking in Model 3 showing much enhanced R-squared, AIC, and BIC. The tree density, transit availability, and crime were found to have significant associations with higher

odds of both walking (OR: 1.05, 1.92, 1.00 respectively) and biking (OR: 1.00, 1.09, 1.00 respectively) compared to in-vehicle status in Model 3. The neighborhood household income, African American population, and traffic crashes were associated with higher odds for walking (OR: 1.00, 1.00, 1.00 respectively), but with lower odds for biking (OR: 0.99, 0.99, 0.99 respectively) compared to in-vehicle. On the other hand, the parks and open spaces were found to be related to lower odds of walking (OR: 0.98) and biking (OR: 0.99) against in-vehicle status.

To determine the effects of all the factors in the probability scale on walking and biking, the average marginal effects were calculated as well. In Model 3, the average marginal effect of transit availability on walking was the highest (0.1061) among all the predictors. It indicated that the probability of walking was approximately 10 % points higher for the areas with great transit accessibility than low levels of transit accessibility. It was also found that when surrounding environments had higher tree density, there was 0.7% higher probability of walking on average than the areas with low tree density.

Table 3.4. Odds ratios and standard errors resulted from Model 1, 2, and 3 to examine the associations between travel modes and tree and park and open space densities, transit availability, incidence of violent crimes, neighborhood American American population and median household income, and traffic crashes involving pedestrians and pedal cyclists adjusted for age, race, household income, and weekday/weekend measured within 200 m buffers of GPS point

Predictors	**Model 1**						**Model 2**						**Model 3**					
	Walking vs. in-vehicle			Biking vs. in-vehicle			Walking vs. in-vehicle			Biking vs. in-vehicle			Walking vs. in-vehicle			Biking vs. in-vehicle		
	OR	95% CI	AME	OR	95% CI	AME	OR	95% CI	AME	OR	95% CI	AME	OR	95% CI	AME	OR	95% CI	AME
Physical environment																		
Greenness - Tree	1.05	(1.04, 1.05)	0.0079	1.01	(1.00, 1.01)	-0.0001	1.05	(1.05, 1.05)	0.0078	1.01	(1.00, 1.01)	-0.0001	1.05	(1.04, 1.05)	0.0077	1.00	(1.00, 1.00)	-0.0000
Land use - Park and open space	0.97	(0.97, 0.97)	-0.0036	1.00	(0.99, 1.00)	0.0001	0.97	(0.97, 0.97)	-0.0036	1.00	(0.99, 1.00)	0.0001	0.98	(0.97, 0.98)	-0.0033	0.99	(0.99, 0.99)	0.0001
Transit availability index	2.06	(2.03, 2.09)	0.1194	1.06	(1.00, 1.12)	-0.0015	2.04	(2.01, 2.07)	0.1173	1.07	(1.05, 1.08)	-0.0014	1.92	(1.92, 1.92)	0.1061	1.09	(1.09, 1.09)	-0.0008
Social environment and safety																		
Crime	1.00	(1.00, 1.00)	0.0001	1.00	(1.00, 1.00)	0.0000	1.00	(1.00, 1.00)	0.0000	1.00	(1.00, 1.00)	0.0000	1.00	(1.00, 1.00)	0.0000	1.00	(1.00, 1.00)	0.0000
Neighborhood median household income													1.00	(1.00, 1.00)	0.0000	0.99	(0.99, 0.99)	-0.0000
Neighborhood African American							1.00	(1.00, 1.00)	0.0000	1.00	(1.00, 1.00)	-0.0000	1.00	(1.00, 1.00)	0.0000	0.99	(0.99, 0.99)	-0.0000
Traffic crash	1.01	(1.01, 1.01)	0.0025	0.99	(0.99, 0.99)	-0.0001	1.01	(1.01, 1.01)	0.0027	0.99	(0.99, 0.99)	-0.0001	1.01	(1.01, 1.01)	0.0021	0.99	(0.99, 0.99)	-0.0001
McFadden's R^2	0.128						0.129						0.140					
Nagelkerke's R^2	0.212						0.212						0.229					
CoxSnell's R^2	0.154						0.155						0.167					
AIC	178352						178285						175644					
BIC	178611						178564						175943					

$p < 0.001$; $p < 0.01$; $p < 0.1$
OR: odds ratio
95% CI: 95% confidence intervals of odds ratios
AME: average marginal effects
AIC: Akaike information criterion
BIC: Bayesian information criterion

3.5 DISCUSSION AND CONCLUSIONS

This study explored how different buffer sizes affected the associations between ATMs and various contexts in the physical and social environment and public safety in the estimation of spatially immediate and temporally momentary exposures accompanying GPS trajectories of individuals for PA and transportation research. It was found that the buffer size had an influence on the associations measured as ORs and the significance levels of variables, and the findings of those two conditions were clearly different in walking and biking. Specifically, the associations of biking and/or walking with parks and open spaces, crime, and traffic crashes did not remain consistent, showing an increase or decrease in ORs moving from positive to negative association or vice versa as the buffer size increased. A possible explanation for the inconsistency is that the changes in direction of the associations between 20 m and 30 m, when compared to other sizes, was caused by the too-small size of 20 m buffer areas, which do not include any park areas, and incidents of crime and traffic crashes around individual paths. Among these three predictors, parks and open spaces particularly showed a decrease in magnitude of its influence on biking over in-vehicle status corresponding to Houston's findings (2014).

Regarding the associations, biking was more sensitive than walking, showing varying statistical significance levels across different buffer sizes over parks and open spaces, transit availability, crime, neighborhood African-American population, and traffic crashes. One common characteristic found in the outcomes was that non-significant associations became significant when the buffer size reached a relatively far distance, like 150 or 200 m. The 200 m distance, particularly in this study, was able to produce more significant variables in Model 3, with better model fits based on several pseudo R-squared measures, than the other two models. Further, Model 3 with 200 m distance showed the best model fits when compared to all the other

shorter buffer distances although this finding was not presented in this study. Neighborhood-level characteristics for demographics and socio-economic status measured dynamically around each buffer of GPS points played an important role in producing the better model, which may be relevant to individuals' perceptions of opportunities for health-promoting behaviors (Boslaugh et al., 2004). Thus, the evidence of the existence of buffer-size effects on multiple types of environmental contexts in this study provided more systematic insights into PA and transportation research with the use of GPS trajectories than previous studies (Rodríguez et al., 2012; Houston, 2014).

In the physical environment, the percentage of tree areas as a proxy of greenness measured within 200 m buffers around each GPS point was one of the consistent predictors, steadily showing significant associations in the three models generated. The tree density showed higher ORs of walking and biking compared to in-vehicle travel. With the objectively measured tree density, this study proved that greenness is likely to be associated with higher walking status compared to in-vehicle, which is consistent with previous studies (Gong et al., 2014; McMorris et al., 2015). The role of parks and open spaces in promoting walking and biking was, however, inconsistent with other studies suggesting that more park areas that individuals are exposed to in their daily trips are significantly associated with the lower OR of active travels, compared to the motorized travel mode (Troped et al., 2003; Sallis et al., 2009; Coombes et al., 2010; Gómez et al., 2010; Boruff et al., 2012; Astell-Burt et al., 2014; Brown et al., 2014; Fisher et al., 2014). One possible explanation is that some adults intentionally take a detour when they drive home, to enjoy the fleeting natural landscape, including green space, which may give in-vehicle status higher odds than two active travel modes (Bell et al., 2015). In addition, since parks and open spaces have more complex aspects, including quality and availability (Lee & Maheswaran,

2011), which may affect associations with ATMs, the findings of parks and open spaces are not as consistent as the tree density. The higher ORs of non-motorized travel modes — walking and biking in this study — against in-vehicle status with regard to transit availability also corresponds to past studies indicating that transit facilities encouraged ATMs and PA (Hoehner et al., 2005; Sallis et al., 2009).

One piece of salient evidence that this study proved was that safety-related factors, including crime and traffic crashes, had significant associations with walking and biking. Compared to traveling by private vehicles or public transit, a high incidence of traffic crashes involving pedestrians and pedal cyclists was significantly associated with the lower likelihood of biking, which provides empirical evidence that traffic crashes constrain PA (Foster & Giles-Corti, 2008). Conversely, there were mixed findings with walking. Unlike biking, walking was found to be more likely to occur in areas with higher traffic crash incidences. Further, the higher incidence of battery was significantly associated with higher ORs of walking and biking compared to in-vehicle. The higher ORs of walking compared to in-vehicle in the associations with crime and traffic crashes were unexpected, suggesting that walking was more likely to happen in areas with more crime cases and traffic crashes in immediate surroundings than in-vehicle status. The reason behind the inconsistent associations could not be clearly explained in this study. However, with the different findings between walking and biking in traffic crashes, this study provided empirical evidence that the mechanism behind the associations between travel modes and some environmental contexts may not work identically, although those are the same ATMs as health-promoting activities. The neighborhood household income and African-American population also supported the disparity in how some environmental contexts could have opposite effects on two different active modes, indicating that walking was likely to be

performed in neighborhoods with a high African-American population and household income compared to in-vehicle status, while biking had opposite outcomes.

The optimized travel mode classification adopted to automatically identify walking, running, biking, and in-vehicle status was one of the innovative parts of this study, and the automatic classification of travel modes only using GPS trajectories showed a remarkable result in accurately identifying those four travel modes. With the newly adopted travel mode classification, this study suggested the novel way of using the predicted travel modes in health, transportation, and urban planning research to understand dynamic exposures to surrounding environments and their impact on individuals' PA, taking into account the daily trips observed by GPS trajectories.

This study, however, has some limitations. First, sampling rate of the GPS trajectories may affect the consistency of the results. In this study, GPS points were sampled at the 10-second interval due to efficient computation of buffer generation; however, coarser (e.g., 60 seconds) or finer (e.g., 5 seconds) sampling scales may have implications of findings related to associations between ATMs and environmental factors depending on different buffer sizes, since the initial analysis with GPS data at the 60-second interval showed somewhat different results in some buffer sizes, yet mostly similar trends in general. Second, the intensity of the travel modes was not considered, which could enrich the understanding of the associations. Walking, especially, can be further separated into light and brisk walking depending on its intensity, which may be differently affected by environmental contexts, as many studies demonstrated by using accelerometer sensors to identify the intensity levels of PA. Last, this study's samples were biased to a specific socio-demographic group, mostly wealthy middle-age whites, which limited

our understanding of how the effects of environmental factors depending on different buffer sizes vary in diverse population groups.

Considering these limitations, future work will further examine the impact of GPS data sampling rates on research findings. The investigation will contribute to mobility research in various fields of study by providing a minimum analytical unit of time for GPS data. In addition, ATMs need to be improved with more categories added, considering the intensity of ATMs, by simply using accelerometer sensors or having them estimated from people's physiological information, including age, height, and weight, and velocity of each walking or biking trip. Moreover, further research will seek to deepen the understanding of the inconsistencies in the results by focusing on different ethnic groups with economic status and gender information. Spatio-temporal analysis will be also useful to look into such inconsistencies in some predictors, like parks and open spaces and safety-related factors in detail, which may be time-sensitive and vary depending on time windows in a day or week and weekend days.

3.6 REFERENCES

Almanza, E., Jerrett, M., Dunton, G., Seto, E. & Pentz, M. A. (2012). A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. *Health & place* **18** (1), 46–54.

Astell-Burt, T., Feng, X. & Kolt, G. S. (2014). Green space is associated with walking and moderate-to-vigorous physical activity (MVPA) in middle-to-older-aged adults: findings from 203 883 Australians in the 45 and Up Study. *British Journal of Sports Medicine* **48** (5), 404–406.

Bell, S. L., Phoenix, C., Lovell, R., & Wheeler, B. W. (2015). Using GPS and geo-narratives: a methodological approach for understanding and situating everyday green space encounters. *Area*, **47** (1), 88–96.

Berke, E. M., Koepsell, T. D., Moudon, A. V., Hoskins, R. E. & Larson, E. B. (2007). Association of the built environment with physical activity and obesity in older persons. *American journal of public health* **97** (3), 486–492.

Berke, E. M., Koepsell, T. D., Moudon, A. V., Hoskins, R. E. & Larson, E. B. (2007). Association of the built environment with physical activity and obesity in older persons. *American journal of public health* **97** (3), 486–492.

Boruff, B. J., Nathan, A. & Nijɲnstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood'in place-based health research. *International journal of health geographics* **11** (1), 22.

Boslaugh, S. E., Luke, D. A., Brownson, R. C., Naleid, K. S. & Kreuter, M. W. (2004). Perceptions of neighborhood environment for physical activity: Is it "who you are" or "where you live?". *Journal of Urban Health* **81** (4), 671–681.

Browning, M. & Lee, K. (2017). Within what distance does “Greenness” best predict physical health? A systematic review of articles with GIS buffer analyses across the lifespan. *International journal of environmental research and public health* **14** (7), 675.

Bureau of Justice Statistics. (2018). Violent crime. Retrieved from https://www.bjs.gov/index.cfm?ty=tp&tid=31

Burgoine, T., Jones, A. P., Brouwer, R. J. N. & Neelon, S. E. B. (2015). Associations between BMI and home, school and route environmental exposures estimated using GPS and GIS: do we see evidence of selective daily mobility bias in children?. *International journal of health geographics* **14** (1), 8.

Cerin, E., Mitős, J., Cain, K. L., Conway, T. L., Adams, M. A., Schofield, G., Sarmiento, O. L., Reis, R. S., Schipperijn, J., Davey, R. & others (2017). Do associations between objectively-assessed physical activity and neighbourhood environment attributes vary by time of the day and day of the week? IPEN adult study. *International Journal of Behavioral Nutrition and Physical Activity* **14** (1), 34.

Chambers, T., Pearson, A., Kawachi, I., Rzotkiewicz, Z., Stanley, J., Smith, M., barr, M., Mhurchu, C. N. & Signal, L. (2017). Kids in space: Measuring children's residential neighborhoods and other destinations using activity space GPS and wearable camera data. *Social Science & Medicine* **193**, 41–50.

Chicago Metropolitan Agency for Planning (CMAP) Data Hub (2017). Transit availability index. Retrieved from https://datahub.cmap.illinois.gov/dataset/access-to-transit-index

Cohen, D. A., Ashwood, J. S., Scott, M. M., Overton, A., Evenson, K. R., Staten, L. K., Porter, D., McKenzie, T. L. & Catellier, D. (2006). Public parks and physical activity among adolescent girls. *Pediatrics* **118** (5), e1381–e1389.

Coombes, E., Jones, A. P. & Hillsdon, M. (2010). The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Social science & medicine* **70** (6), 816–822.

Dunton, G. F., Almanza, E., Jerrett, M., Wolch, J. & Pentz, M. A. (2014). Neighborhood park use by children: use of accelerometry and global positioning systems. *American journal of preventive medicine* **46** (2), 136–142.

Fisher, K. J., Li, F., Michael, Y. & Cleveland, M. (2004). Neighborhood-Level Influences on Physical Activity among Older Adults: A Multilevel Analysis. *Journal of Aging and Physical Activity* **12** (1), 45-63.

Foster, S. & Giles-Corti, B. (2008). The built environment, neighborhood crime and constrained physical activity: An exploration of inconsistent findings. *Preventive Medicine* **47** (3), 241–251.

Gómez, L. F., Parra, D. C., Buchner, D., Brownson, R. C., Sarmiento, O. L., Pinzón, J. D., Ardila, M., Moreno, J., Serrato, M. & Lobelo, F. (2010). Built Environment Attributes and Walking Patterns Among the Elderly Population in Bogotá. *American Journal of Preventive Medicine* **38** (6), 592–599.

Gong, Y., Gallacher, J., Palmer, S. & Fone, D. (2014). Neighbourhood green space, physical function and participation in physical activities among elderly men: the Caerphilly Prospective study. *International Journal of Behavioral Nutrition and Physical Activity* **11** (1), 40.

Harrison, F., Burgoine, T., Corder, K., van Sluijs, E. M. & Jones, A. (2014). How well do modelled routes to school record the environments children are exposed to?: a cross-sectional comparison of GIS-modelled and GPS-measured routes to school. *International*

journal of health geographics **13** (1), 5.

Hillsdon, M., Panter, J., Foster, C. & Jones, A. (2006). The relationship between access and quality of urban green space with population physical activity. *Public health* **120** (12), 1127–1132.

Hirsch, J. A., Winters, M., Ashe, M. C., Clarke, P. J. & McKay, H. A. (2016). Destinations That Older Adults Experience Within Their GPS Activity Spaces: Relation to Objectively Measured Physical Activity. *Environment and Behavior* **48** (1), 55–77.

Hirsch, J. A., Winters, M., Clarke, P. & McKay, H. (2014). Generating GPS activity spaces that shed light upon the mobility habits of older adults: a descriptive analysis. *International journal of health geographics* **13** (1), 51.

Hoehner, C. M., Ramirez, L. K. B., Elliott, M. B., Handy, S. L. & Brownson, R. C. (2005). Perceived and objective environmental measures and physical activity among urban adults. *American Journal of Preventive Medicine* **28** (2, Supplement 2), 105–116.

Houston, D. (2014). Implications of the modifiable areal unit problem for assessing built environment correlates of moderate and vigorous physical activity. *Applied Geography* **50**, 40–47.

James, P., Berrigan, D., Hart, J. E., Hipp, J. A., Hoehner, C. M., Kerr, J., Major, J. M., Oka, M. & Laden, F. (2014). Effects of buffer size and shape on associations between the built environment and energy balance. *Health & place* **27**, 162–170.

Jankowska, M. M., Natarajan, L., Godbole, S., Meseck, K., Sears, D. D., Patterson, R. E. & Kerr, J. (2017). Kernel Density Estimation as a Measure of Environmental Exposure Related to Insulin Resistance in Breast Cancer Survivors. *Cancer Epidemiology and Prevention Biomarkers*.

Kwan, M.-P. (1999). Gender and individual access to urban opportunities: a study using space–time measures. *The Professional Geographer* **51** (2), 210–227.

Lee, A. C., & Maheswaran, R. (2011). The health benefits of urban green spaces: a review of the evidence. *Journal of public health*, **33** (2), 212–222.

Lee, N. C., Voss, C., Frazer, A. D., Hirsch, J. A., McKay, H. A. & Winters, M. (2016). Does activity space size influence physical activity levels of adolescents?—A GPS study of an urban environment. *Preventive Medicine Reports* **3**, 75–78.

Maas, J., Verheij, R. A., Spreeuwenberg, P. & Groenewegen, P. P. (2008). Physical activity as a possible mechanism behind the relationship between green space and health: a multilevel analysis. *BMC public health* **8** (1), 206.

McGinn, A. P., Evenson, K. R., Herring, A. H., Huston, S. L. & Rodriguez, D. A. (2007). Exploring associations between physical activity and perceived and objective measures of the built environment. *Journal of Urban Health* **84** (2), 162–184.

Mitchell, C. A., Clark, A. F. & Gilliland, J. A. (2016). Built Environment Influences of Children's Physical Activity: Examining Differences by Neighbourhood Size and Sex. *International Journal of Environmental Research and Public Health* **13** (1).

Nagel, C. L., Carlson, N. E., Bosworth, M. & Michael, Y. L. (2008). The relation between neighborhood built environment and walking activity among older adults. *American journal of epidemiology* **168** (4), 461–468.

National Institute of Justice. (2018). Violent crimes. Retrieved from https://www.nij.gov/topics/crime/violent/Pages/welcome.aspx

Perchoux, C., Kestens, Y., Thomas, F., Hulst, A. V., Thierry, B. & Chaix, B. (2014). Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and

associations with transportation modes (the RECORD Cohort Study). *Social Science & Medicine* **119**, 64–73.

Prins, R. G., Pierik, F., Etman, A., Sterkenburg, R. P., Kamphuis, C. & Van Lenthe, F. (2014). How many walking and cycling trips made by elderly are beyond commonly used buffer sizes: results from a GPS study. *Health & place* **27**, 127–133.

Rodriguez, D. A., Cho, G.-H., Evenson, K. R., Conway, T. L., Cohen, D., Ghosh-Dastidar, B., Pickrel, J. L., Veblen-Mortenson, S. & Lytle, L. A. (2012). Out and about: association of the built environment with physical activity behaviors of adolescent females. *Health & place* **18** (1), 55–62.

Rundle, A. G., Sheehan, D. M., Quinn, J. W., Bartley, K., Eisenhower, D., Bader, M. M., Lovasi, G. S. & Neckerman, K. M. (2016). Using GPS data to study neighborhood walkability and physical activity. *American journal of preventive medicine* **50** (3), e65–e72.

Sallis, J. F., Bowles, H. R., Bauman, A., Ainsworth, B. E., Bull, F. C., Craig, C. L., Sjöström, M., Bourdeaudhuij, I. D., Lefevre, J., Matsudo, V., Matsudo, S., Macfarlane, D. J., Gomez, L. F., Inoue, S., Murase, N., Volbekiene, V., McLean, G., Carr, H., Heggebo, L. K., Tomten, H. & Bergman, P. (2009). Neighborhood Environments and Physical Activity Among Adults in 11 Countries. *American Journal of Preventive Medicine* **36** (6), 484–490.

Schipperijn, J., Bentsen, P., Troelsen, J., Toftager, M. & Stigsdotter, U. K. (2013). Associations between physical activity and characteristics of urban green space. *Urban Forestry & Urban Greening* **12** (1), 109–116.

Stewart, C. A., Cockerill, T. M., Foster, I., Hancock, D., Merchant, N., Skidmore, E., Stanzione, D., Taylor, J., Tuecke, S., Turner, G., Vaughn, M. & Gaffney, N. I. (2015). Jetstream: A

Self-provisioned, Scalable Science and Engineering Cloud Environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* (pp. 29:1–29:8).

Thierry, B., Chaix, B. & Kestens, Y. (2013). Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International Journal of Health Geographics* **12** (1), 14.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R. & Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering* **16** (5), 62–74.

Troped, P. J., Saunders, R. P., Pate, R. R., Reininger, B. & Addy, C. L. (2003). Correlates of recreational and transportation physical activity among adults in a New England community. *Preventive Medicine* **37** (4), 304–310.

United States Census Bureau (2010). 2010 Census summary of Chicago city, Illinois. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

Wang, J. & Kwan, M.-P. (2018). An Analytical Framework for Integrating the Spatiotemporal Dynamics of Environmental Context and Individual Mobility in Exposure Assessment: A Study on the Relationship between Food Environment Exposures and Body Weight. *International Journal of Environmental Research and Public Health* **15** (9).

Winters, M., Brauer, M., Setton, E. M. & Teschke, K. (2010). Built environment influences on healthy transportation choices: bicycling versus driving. *Journal of urban health* **87** (6), 969–993.

Yin, L., Raja, S., Li, X., Lai, Y., Epstein, L. & Roemmich, J. (2013). Neighbourhood for

Playing: Using GPS, GIS and Accelerometry to Delineate Areas within which Youth are Physically Active. *Urban Studies* **50** (14), 2922–2939.

Zenk, S. N., Schulz, A. J., Matthews, S. A., Odoms-Young, A., Wilbur, J., Wegrzyn, L., Gibbs, K., Braunschweig, C. & Stokes, C. (2011). Activity space environment and dietary and physical activity behaviors: A pilot study. *Health & Place* **17** (5), 1150–1161.

CHAPTER 4: INTERPRETATION OF CONTEXTUAL INFLUENCES WITH MACHINE LEARNING TECHNIQUES: TRAVEL MODE LIKELIHOOD MAPPING USING GPS TRAJECTORIES OF INDIVIDUALS

4.1 INTRODUCTION

In the last decade or so, a number of scholars have examined the health impact of various environmental factors (e.g. the built environment) on moderate to vigorous physical activity (PA), like jogging and brisk walking (Physical Activity Guidelines Advisory Committee, 2008). They have conducted studies that have taken into account the daily movements of individuals using GPS and accelerometers (Cooper et al., 2010; Troped et al., 2010; Almanza et al., 2012; Boruff, Nathan & Nijënstein, 2012; Lachowycz et al., 2012; Rodríguez et al., 2012; Jansen et al., 2016). The rationale behind the use of GPS trajectories comes from the important role of non-residential environments or contexts in people's daily lives in addition to their residential neighborhoods (Roux & Mair, 2010; Perchoux et al., 2013). Kwan (2012, 2013, 2018b) has articulated this issue as the uncertain geographic context problem (UGCoP), which refers to the notion that the effects of area-based attributes (e.g., land-use mix) on individual behaviors or outcomes (e.g., PA) can be affected by the way in which contextual units or neighborhoods are geographically delineated. It "arises because of the spatial uncertainty in the actual areas that exert contextual influences on the individuals being studied and the temporal uncertainty in the timing and duration in which individuals experienced these contextual influences" (Kwan 2012, p. 959).

Past studies on the impact of environmental contexts on active travel modes (ATMs) as a subset of PA, however, have often yielded inconsistent results (Sallis, Bauman & Pratt, 1998; King et al., 2002; Spence & Lee, 2003; McNeill, Kreuter & Subramanian., 2006; Sallis et al.,

2012; Sallis, Owen, & Fisher, 2015). Thus, they have failed to provide reliable empirical evidence that enables the production of generalizable knowledge. The mixed associations between environmental contexts and ATMs in past studies were observed to identify various environmental factors, including greenness (Troped et al., 2010), recreational facilities (Hoehner et al., 2005), traffic volumes (McGinn et al., 2007; Nagel et al., 2008), and the density of road intersections (Troped et al., 2010). To address the abovementioned inconsistency, this study considers three methodological issues. First, since the interactions between human behaviors and environments are complex, a lot of physical and social environmental factors need to be considered with respect to any modeling effort. Traditional statistical methods, like regression models, are not suitable for conducting data-driven analyses that involve many variables and the task of finding meaningful patterns using a massive amount of GPS points. Classical methods, like logistic regression models, fundamentally require interactions of interest to be pre-specified, which do not allow for the investigation of complex interactions between many predictors and outcome. In addition, van der Ploeg and colleagues (2014) claimed that modern modeling methods, such as machine learning algorithms, are more appropriate for a large amount of data points than classical modeling methods. Second, the association between a particular environmental factor and an ATM may vary over space and, thus, cannot be generalized at a global level. This phenomenon is referred to as spatial non-stationarity, a situation in which the effect of a contextual factor on or its association with a particular health outcome may vary across different geographic locations. Further investigations on a local level at a finer spatial scale are, thus, often needed as a key that can address spatial non-stationary and the spatially varying local impact of specific environmental factors (Wang, Lee & Kwan, 2018). Third, in addition to the geographic context, the temporal context might also account for part of the

inconsistency in the findings of past studies. The role of time in estimating dynamic environmental exposures is especially critical to understanding the neighborhood effect on human behaviors (Kwan, 2013, 2018a; Park & Kwan, 2017; Wang & Kwan, 2018a). For example, the effects of certain environmental factors may vary temporally and become significant during a particular time period in a day (Cerin et al., 2017).

To address these three methodological issues, this study adopts machine learning techniques to create likelihood maps of ATMs. Likelihood maps have been widely used in epidemiology and public health to predict the likeliness of risks of infection from viruses or diseases through spatial analysis or statistical analysis/modeling (Miranda et al., 2002; Ruiz et al., 2004; Chang et al., 2009; Keddem et al., 2015). With the advancement in artificial intelligence techniques, likelihood mapping has recently deployed machine learning models in various research domains to construct predictive models for monitoring geological hazards and natural disasters and detecting natural resources by predicting the likelihood of various phenomena or their probable locations (Ruiz et al., 2010; Pradhan, 2013; Tehrany et al., 2014; Naghibi et al., 2016). Predictive models can capture complex interactions between such phenomena and many environmental factors and forecast events over a study area including 'blind spots.' For instance, considering GPS trajectories as observations, blind spots are areas without GPS trajectories and, thus, have no identified patterns due to the lack of data points relating to the use of conventional visualization methods (e.g. kernel density estimation). Particularly, GPS trajectories are not static observations but dynamic movements. Thus, predicting and visualizing ATM over a whole city, which aid government policies and interventions to understand people's mobility and promote people's health, are challenging. Further, machine learning especially has evolved to facilitate the construction of predictive

models in an interpretable form (Letham et al., 2015; Krause et al., 2016). In other words, there is a new path that can lead to a more in-depth understanding of human behaviors and environments which has been opened up through the use of machine learning techniques. Therefore, in this study, the prediction capabilities of machine learning enable us to identify travel modes in all neighborhoods of an entire city, and machine learning is one key technique that scrutinizes the complex associations between travel modes and multiple environmental variables, while considering the spatially and temporally dynamic characteristics of GPS trajectories and interprets the global and local impact of environmental factors on travel modes. As Dodge (2016) suggested, the potential of mobility-related data is not limited to shaping knowledge acquaintance from movement records associated with given contexts but also can be extended to the prediction of behaviors under certain environmental conditions.

By utilizing machine learning techniques, this study contributes to the development of a data-driven approach to research concerning public health and transport. In addition, the proposed approach for generating likelihood maps will advance GIS mapping by employing machine learning techniques to predict spatio-temporal patterns regarding public health, using the daily movement data of individuals. In this regard, this study facilitates the introduction of predictive power and abilities to GIScience for interpreting complex interactions between human behaviors and environments.

The remainder of the paper is organized as follows. We discuss previous research on likelihood mapping in Section 4.2. In Section 4.3, a proposed framework termed as Spatio-temporal Mapping And INterpretation (SMAIN) is introduced, and the geospatial data gathered for a case study are described. Then, in Section 4.4, SMAIN is applied to explore the global and local impact of certain geographic contexts on travel modes in Chicago, U.S. Section 4.5

concludes by providing potential ways of using the suggested framework and discusses the limitations and possible future work.

4.2 LIKELIHOOD MAPPING

Likelihood mapping has been used in a variety of disciplines, including epidemiology and public health, to predict the probable risks of infection from viruses or disease (e.g. West Nile virus outbreak or lead poisoning) through spatial analysis or statistical analysis/modeling primarily at the level of administrative boundaries or census boundaries or tax parcels (Miranda et al., 2002; Ruiz et al., 2004; Keddem et al., 2015). The utilization of machine learning models have increased for likelihood mapping in the fields, like natural disasters and natural resource management, due to the advancement in artificial intelligence techniques and the increasing demand for understanding the complex relationships among environmental factors (Pradhan, 2013; Tehrany et al., 2014; Naghibi et al., 2016; Song et al., 2017).

Likelihood maps have been developed to visually represent not only the likelihood of exposures to harmful environments that may influence people but also the likelihood of geological hazards and natural disasters as well as the probable locations of natural resources. With regard to environmental health problems, Miranda et al. (2002) created a map to examine the extent to which children have a high risk of lead poisoning in Durham in North Carolina, U.S. Different levels of lead-related risks were predicted using multivariate statistical analysis and mapped at the tax parcel level. Keddem et al. (2015) applied mixed methods to identify the environmental triggers of asthma and evaluate risks of asthma control through a map that considered pertinent environmental factors. To determine the relevant environmental characteristics, qualitative interviews were conducted in West Philadelphia, and the perceived risk to asthma control was estimated by summing up the standardized scores of the factors and

visualizing the results on a map. Besides the likelihood of exposure to adverse environmental conditions, the risks of infectious diseases have been explored through spatial mapping. The risk of contracting Dengue fever in Jalore, India, was predicted using a regression model that incorporated various sociocultural variables and mapped administrative areas with five risk levels (Bohra & Andrianasolo, 2001). Risk maps of the West Nile virus, also spread by mosquito bites in a manner similar to dengue fever, in and around the Chicago area were generated based on logistic regression models that took into account various factors to identify the spatial patterns of high-risk areas (Ruiz et al., 2004).

Good examples of likelihood maps for geological hazards and natural disasters are landslide susceptibility (Pradhan, 2013; Bui et al., 2016a), flood (Tehrany et al., 2014), and fire occurrence (Song et al., 2017; Bui et al., 2016b). GIS-based likelihood mapping is also an advanced visualization method used to find possible areas of groundwater springs as a natural resource that can help address the problem of insufficient water supply for agricultural, industrial, and domestic uses (Naghibi et al., 2016; Zabihi et al., 2016; Lee et al., 2017). In these studies, machine learning models comprised the methodological foundation for building models and predicting phenomena of interest and use training data extracted from relevant environmental GIS data. The resultant maps were visualized through raster data layers, where the value of each pixel indicates different levels of susceptibility.

Likelihood mapping has not been applied to mobility research despite its huge potential for examining individuals' movements using different travel modes and its association with specific characteristics of the environment. Particularly, mapping travel modes have been limited to the simple visualization of commuting routes or the spatiotemporal visualization of travel choice using an interpolation method on point observations without considering contextual

influences (Cooper et al., 2010; Delmelle & Delmelle, 2012). In this study, the likelihood of the occurrence of different travel modes in Chicago is mapped based on machine learning models which are trained using predictors extracted from various physical and social environmental factors. A high likelihood of the occurrence of a specific travel mode (hereafter referred to as travel mode occurrence) in this study refers to the travel mode which is found to frequently appear at a certain place. Likelihood maps can retain and exhibit the spatiotemporal characteristics of different travel modes across a study area with the help of the predictive power of machine learning. The recent improvements in machine learning techniques regarding the interpretation of predictions of trained models also help to account for any difference that may exist between global and local impact of environmental factors on travel modes. Here, the likelihood mapping of travel mode occurrence is different from the method used in previous studies in that this study adopts the dynamic individuals' movements represented by GPS trajectories to detect potential areas of travel modes on a map rather than fixed geographic locations that represent some events or area-based units, like administrative or census boundaries.

4.3 METHOD

4.3.1 SMAIN - A framework for generating and interpreting likelihood maps

A framework of the whole process for generating likelihood maps based on GPS trajectories and interpreting these maps is formalized as illustrated in Figure 1. The formalized structure involves data input, GIS analysis, the tasks of training and testing machine learning models to create likelihood maps, and global and local interpretation, which can be applied to any mobility research using GPS trajectories. In the training process, the GPS data with labels

(indicating travel modes, such as walking, biking, and in-vehicle status in this study) go through a feature-extraction process meant to create multiple variables for environmental factors. As measurable characteristics, features are one of the fundamental elements in machine learning used for training a model to predict a phenomenon with continuous values or different categories (e.g. the percentage of trees in this study). In this framework, features that represent each type of environmental factors or geographic contexts are extracted using geospatial data. To estimate individual exposures to these factors or contexts, a delineating method (e.g. buffer analysis) is used to define the contextual boundaries based on GPS points. Specifically, 34 features are derived from multiple physical and social contexts in this study. The extracted features are then used to train predictive models to produce the likelihood of travel mode occurrence in this study.

In order to choose the best predictive model, different kinds of machine learning models are tested and their performances are compared. The trained models are then applied to unlabeled data employing the same predictors which have been used in the training process. The features of the unlabeled data are calculated for every point that is regularly distributed, which are assumed together as simulated GPS points, across a study area. Since the goal of likelihood mapping is to predict and explore spatial patterns over the entire study area, the same features used in the training process are extracted from the characteristics of the surrounding environments of all the regularly distributed points to predict and eventually map the predicted travel modes in this study. Trained models are quantitatively and qualitatively evaluated to choose the best model for drawing likelihood maps. Ten-fold cross-validation with confusion matrices is one of the methods used for evaluating the performance of the models in a quantitative fashion. The predicted patterns of different labels assigned to the regularly distributed points are visualized on a map, and their spatial patterns are qualitatively examined. Once the evaluation process is

performed, the best machine learning model is chosen, and likelihood maps are created based on the selected model to explore and investigate the predicted patterns of various labels (e.g. walking, in-vehicle, biking) at different time points (e.g. weekdays versus weekend days). The chosen model and the produced maps work collectively to compare the relative global and local importance of environmental factors and investigate the associations between travel modes and these factors using explanatory tools. In this study, the global and local impact are illustrated by partial dependence and centered individual conditional expectation (ICE) plots (Friedman, 2001; Goldstein, 2015). The marginal effect of environmental factor(s) of interest on the prediction of the global model is observed at the average and individual levels using the two aforementioned plots. Regarding local understanding, local interpretable model-agnostic explanations (LIME) are helpful for explaining every prediction of individual observations based on the assumption that complex machine learning models can be locally accounted with linear models (Ribeiro et al., 2016). LIME generate feature weights in order to provide explanations of the local behavior of important environmental factors.

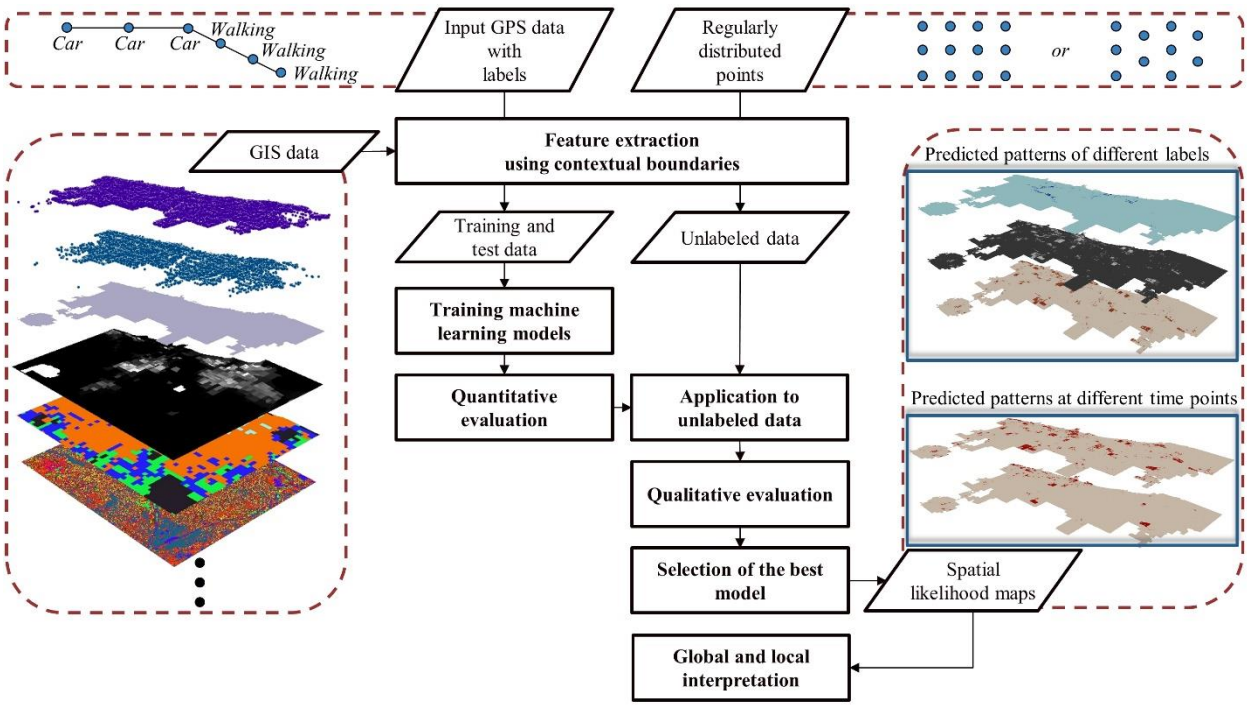


Figure 4.1. The Spatio-temporal Mapping And INterpretation (SMAIN) framework for creating likelihood maps based on GPS trajectories and for interpreting generated maps. Input data: individuals' GPS data for training models and regularly distributed points over a research area to be predicted as travel modes using the trained models. Feature extraction: variable computation from multiple environmental contexts based on pre-defined exposure areas along GPS trajectories. Evaluation: quantitative and visual validations of the performance of trained classifiers. Interpretation: examination of the learned interactions between predicted human behaviors and environmental contexts in global and local models using explanatory tools.

4.3.2 Case study

This study seeks to classify travel modes using GPS trajectories and create its likelihood maps in Chicago. GPS trajectory data of 168 adults obtained from the Chicago Regional Household Travel Inventory (CRHTI) project, described in Section 3.3.1, are used in this study to identify the travel modes present in Chicago and explore their spatiotemporal patterns, which are determined by environmental factors on likelihood maps. To obtain the travel modes of trips, the same classification algorithm developed in Section 3.3.2 was applied.

4.3.3 Likelihood mapping with environmental factors

Buffers are used to define the contextual areas that samples were exposed to along the routes of their daily trips. Specifically, 200 m circular buffers are created around each GPS point to calculate the features. The 200 m distance was found to adequately catch the characteristics of various environments in Chapter 3. As shown in Table 4.1, the predictors are 35 measures calculated to represent various physical, social, and safety-related environmental categories. Eleven types of violent crimes are separately considered: assault, battery, burglary, kidnapping, offense involving children, other offense, robbery, sex offense, sexual assault, stalking, and weapon violation (Bureau of Justice Statistics, 2018; National Institute of Justice, 2018). Six types of land use are included: commercial area, industrial area, institutional area, residential area, parks and open spaces, and religious place. The predictor for safety-related city services includes the complaints regarding abandoned vehicles and graffiti. The two separate types of traffic crashes refer to collisions involving pedestrians or pedal cyclists. I tried to gather and use data as close to the year 2007 as possible, since that was the year in which the GPS data were collected.

Table 4.1. Description of predictors

Category	Data source	Measure	Measuring unit	Resolution/ unit	Time	Comments
Bare soil	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	
Building	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	

Table 4.1 (cont.)

Category	Data source	Measure	Measuring unit	Resolution/ unit	Time	Comments
Crime	Chicago Data Portal	Count	Number of crimes/km^2	Point	2007	11 predictors - assault, battery, burglary, kidnapping, offense involving children, other offense, robbery, sex offense, sexual assault, stalking, weapon violation
Greenness	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	2 predictors - grass, tree
Land use	Chicago Data Portal	Percentage	Area (m^2)	Polygon	2010	6 predictors - commercial area, industrial area, institutional area, residential area, park and open space, religious place
Neighborhood household income	American Community Survey of United States Census Bureau	Average	Dollars ($)	Census tract (polygon)	2010	
Neighborhood ethnic group	American Community Survey of United States Census Bureau	Population	Number of people/km^2	Census tract (polygon)	2010	4 predictors – African American, white, Hispanic, Asian
Paved surface	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	Other paved surfaces except for road

Table 4.1 (cont.)

Category	Data source	Measure	Measuring unit	Resolution/ unit	Time	Comments
Road	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	
Safety-related city services	Chicago Data Portal	Count	Number of reports/km^2	Point	2010	Complaints about abandoned vehicles and graffiti
Transit availability index	Chicago Metropolitan Agency for Planning Data Hub	Average	Access to Transit Index	Polygon	2010	
Traffic crash	Illinois Department of Transportation	Count	Number of crashes/km^2	Point	2007	2 predictors - pedestrian, pedal cyclist
Walkable and bikeable paths	OpenStreetMap	Length	Length (m/m^2)	Road segment (polyline)	2010	2 predictors - walkable, bikeable
Water	Land Cover data from Chicago Metropolitan Agency for Planning Data Hub	Percentage	Area (m^2)	1 m pixel	2010	

Three different supervised machine learning algorithms — support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGB) — are trained to predict travel modes and produce likelihood maps. Specifically, for the kernel function of SVM, the radial basis is selected since it has shown better performance compared to other kernel types (not presented here). XGB is an implementation of gradient boosting for more efficient computation and scalability and for mitigation of over-fitting (Chen & Guestrin, 2016). RF and XGB particularly provide an indicator to identify variables having high importance. Using this measure, we can identify the environmental factors which greatly affect ATMs. The performance

and the predicted results of SVM, RF and XGB are compared and discussed. The predictive accuracy of the three machine learning algorithms is summarized in a confusion matrix. The performance of the three models is validated by using 10-fold cross-validation. Overall accuracy calculated through the validation shows the machine learning algorithm which predicts travel modes more accurately than the others. R statistical software is used to run and evaluate the predictive models. XSEDE Jetstream with Intel Xeon E-2680v3 CPUs (24 cores) aids accelerating running feature calculation and training machine learning models (Towns et al., 2014; Stewart et al., 2015)

To apply the generated predictive models to the entire Chicago area, regularly distributed points across this area are generated, and the same features used in the three models are also calculated for all the sample points to predict travel modes all over the area while taking into account the surrounding environments. The likelihood of the occurrence of walking, biking, and traveling in a vehicle is represented on the hexagonal grid. A hexagon is the most complex regular polygon, and contiguous hexagons can fill a plane without leaving gaps, unlike the rectangular grid. Contrary to the rectangular grid on the left in Figure 4.2, the distance between one centroid and any of the six neighboring centroids is the same. The equal distance to the cell centroids of the neighboring cells especially works as a salient condition to reduce sampling bias of points (Birch et al., 2007; Wang & Kwan, 2018b). This means that the hexagonal grid generates a more accurate sampling and is more isotropic in terms of its six directions to neighboring hexagons. The distance between two centroids (d) is set to 30 m in this study. Then, a larger-sized hexagon mesh (d = 100 m) is created for the purpose of visualization. It aggregates the predicted travel modes of the regularly distributed points and calculates the likelihood of occurrence of each travel mode on the respective hexagon.

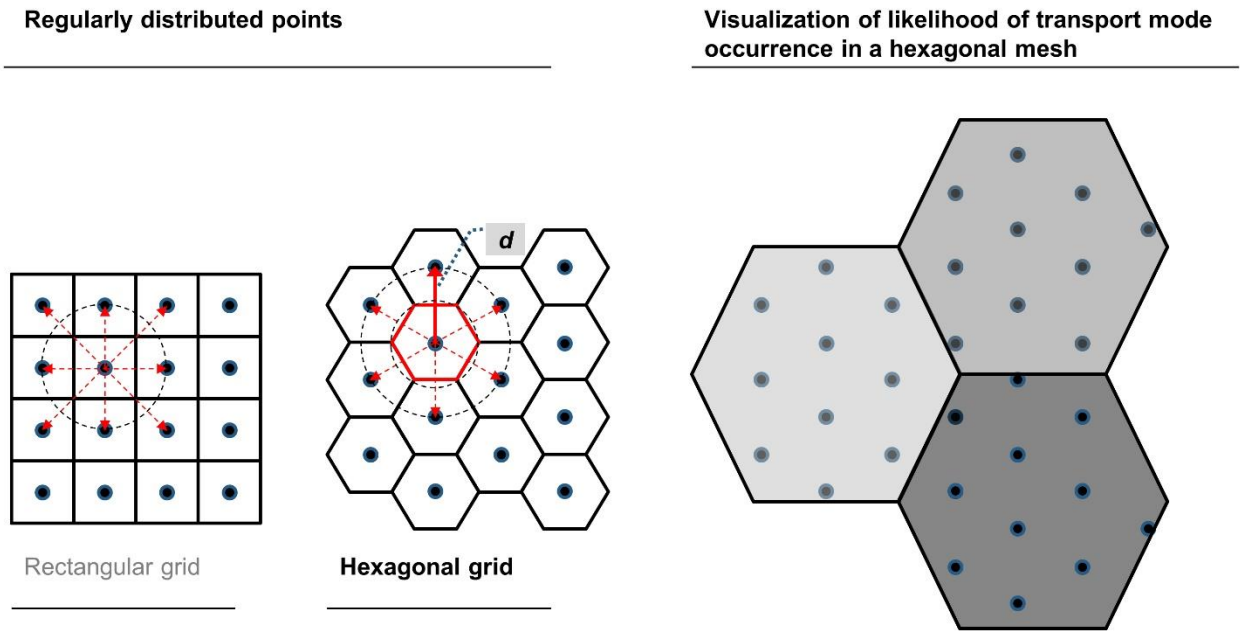


Figure 4.2. Generation of regularly distributed points and visualization of the likelihood of travel mode occurrence

The visual exploratory analysis and interpretation of the associations between specific travel modes and environmental contexts are performed using likelihood maps created by the best model, which yields the highest performance in the quantitative evaluation. The predicted likelihood maps are explored to understand the broad ATM patterns in two different Chicago communities — Lakeview and Chatham. Lakeview is a lively neighborhood in Chicago with a variety of commercial places and recreational areas. Most of its residents are whites with high per capita income (Chicago Data Portal, 2013), and young adults constitute the dominant age group. Chatham is a community which has the highest rates of crime in Chicago (Chicago Police Department, 2007). The majority of the population in Chatham consists middle-class African Americans with lower per capita income ($18,881) than Lakeview (Chicago Data Portal, 2013).

In addition, ATM patterns during weekdays or weekends are examined through the resulted maps. Specifically, to identify the role of crime in ATMs at specific time points during a

day on weekdays or weekends, the variations of some crime factors in their relative importance in predicting walking are explored along with graphs and partial dependence plots with ICE curves. Separate models are trained to predict such weekdays' and weekends' travel mode patterns and even at different time points during a day.

4.4 RESULT

4.4.1 Quantitative and qualitative evaluation of radial SVM, RF, and XGB

The total number of GPS points available as observations was 1,781,710, with 35 features characterizing the surrounding environments along the GPS trajectories of the 168 adults. The performance of the trained models of radial SVM, RF, and XGB was quantitatively evaluated using 10-fold cross-validation. With regard to the number of instances, 68,715 observations were randomly selected from the total number of observations for more efficient computation. Since the percentage of the biking mode was too low (2%) among the three modes, the Synthetic minority over-sampling technique method (Chawla et al., 2002) was used to address the problem of class imbalance by generating more observations for biking. All three machine learning models were tuned with hyperparameters through grid searches. Overall, the best-trained models of RF and XGB showed the highest predictive accuracy — 87.81 and 87.82 % respectively — when optimal parameters were applied (Table 4.2 and Figure 4.3). Even though the RF classifier showed the slightly lower overall accuracy, walking and in-vehicle status were classified more accurately than XGB as shown in Table 4.2. Further, RF was mostly superior to SVM and XGB in not only its sensitivity but also its specificity, and its positive and negative predictive values (Table 4.3).

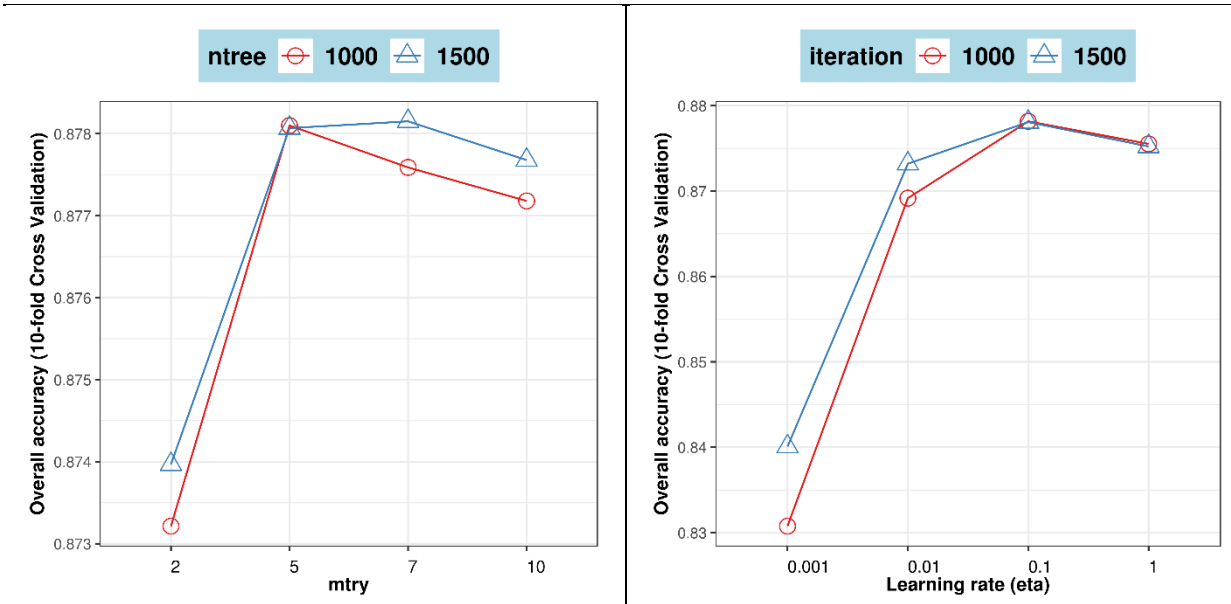
Table 4.2. The overall predictive accuracy of radial support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGB) classifiers to classify walking, in-vehicle, and biking modes

		Classifier		
		Radial SVM	**RF**	**XGB**
Travel mode	**Walking (%)**	83.34	87.09	86.63
	In-vehicle (%)	86.68	88.30	88.55
	Biking (%)	76.03	78.15	74.19
Overall accuracy (%)		85.63	87.81	87.82
Kappa		0.678	0.725	0.723

Table 4.3. Sensitivity, specificity, and positive and negative predictive values of radial support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGB)

	Radial SVM			**RF**			**XGB**		
	W^1	**I^2**	**B^3**	**W^1**	**I^2**	**B^3**	**W^1**	**I^2**	**B^3**
Sensitivity	83.34	86.68	76.03	87.09	88.30	78.15	86.63	88.55	74.19
Specificity	92.88	85.31	94.89	93.20	88.33	96.02	93.02	87.49	96.39
Positive predictive value	80.58	93.88	19.73	81.95	95.16	24.47	81.48	94.84	25.33
Negative predictive value	94.02	71.14	99.59	95.32	74.40	99.63	95.15	74.64	99.56

* W^1: Walking, I^2: In-vehicle status, B^3: Biking



(a) Random forest tuning parameters

(b) Extreme gradient boosting tuning parameters

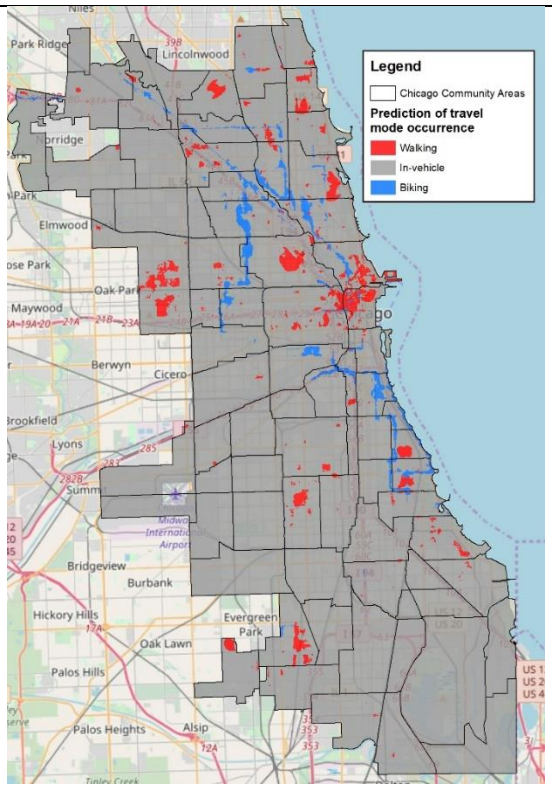
Figure 4.3. Random forest and extreme gradient boosting classifiers with tested tuning hyperparameters (ntree: number of trees). (a) variation of the overall accuracy across different values for the number of variables randomly sampled at each split (mtry), (b) variation of the overall accuracy across different values for the learning rate (eta).

Considering the best RF model with relatively high performance, a qualitative evaluation was conducted by producing likelihood maps using the three classifiers (Figure 4.4). Airports were excluded in the prediction since those areas had no value for median household income. Both RF and XGB portrayed similar travel mode patterns across the Chicago area, not only in places where GPS trajectories existed, but also in areas where GPS trajectories were not found. SVM, however, showed dispersed and relatively weak patterns of walking and biking compared to RF and XGB, which did not help narrow down specific places for discovering the common characteristics of travel modes associated with different environments. Thus, based on the quantitative and qualitative evaluation, RF was selected as the best model to create further likelihood maps.

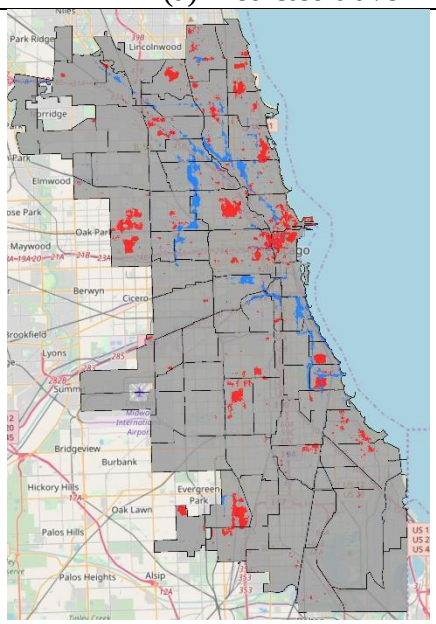
4.4.2 Global impact of environmental contexts on travel modes

As shown in Figure 4.5, travel mode likelihood maps were created using RF. Maps were drawn with either the likelihood of the predicted occurrence of each travel mode (Figure 4.5-(b), (c)) or the predicted occurrence of major travel modes, by showing a travel mode on each hexagon (d = 100 m) with the highest likelihood among all the three travel modes (Figure 4.5-(a)). In Chicago, most areas were predicted to have "traveling in a vehicle" as the dominant travel mode, whereas walking and biking were found over a wide area including the downtown of Chicago. Especially, people were found to perform high levels of walking and biking in recreational and commercial areas in and around downtown Chicago (Figure 4.5-(d), (e)).

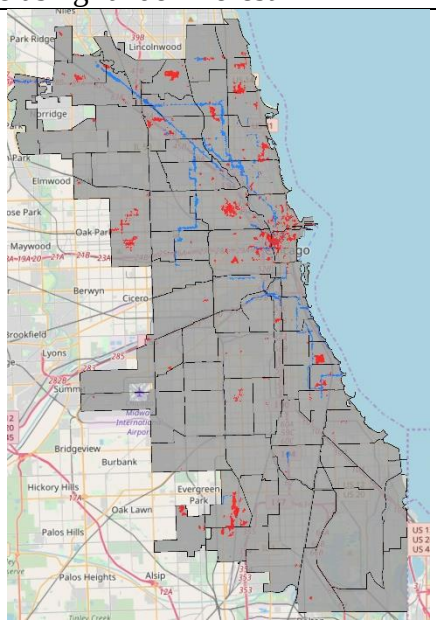
With regard to the relative importance of features in the RF model, different ethnic group populations (African American, Hispanic, White, Asian), built environment (Tree, Bikeable and Walkable paths, Pavement, Grass, Road, Building) neighborhood income (Income), residential (ResidLU) and commercial (CommLU) land uses, crime of battery, and parks and open spaces were found to have relatively high global importance in association with physical and social contexts (Figure 4.6). Especially, Tree was the most influential feature regarding greenness showing the highest importance when predicting walking. Among 11 violent crimes, battery showed relatively high importance.



(a) Predicted travel mode occurrence using random forest



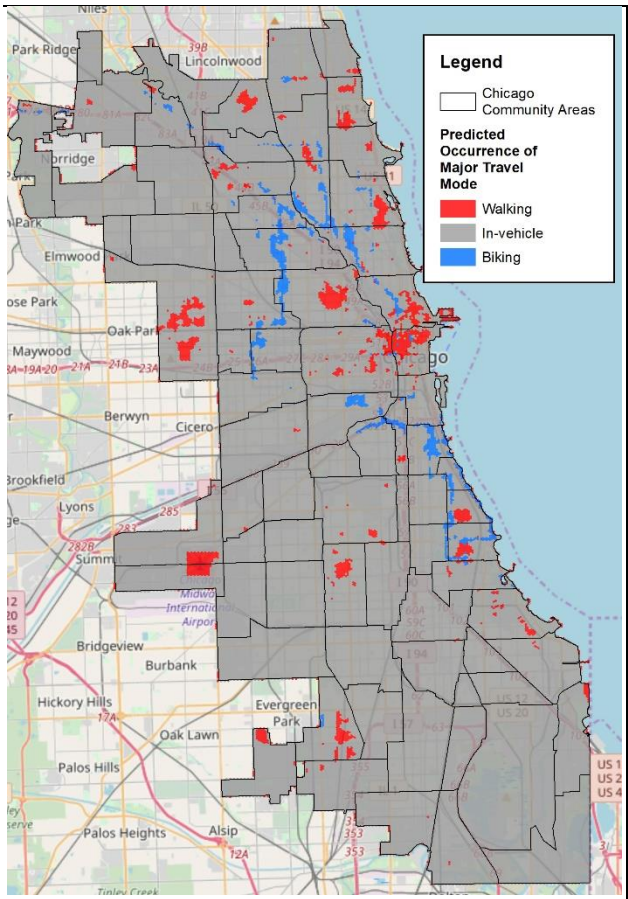
(b) Predicted travel mode occurrence using extreme gradient boosting



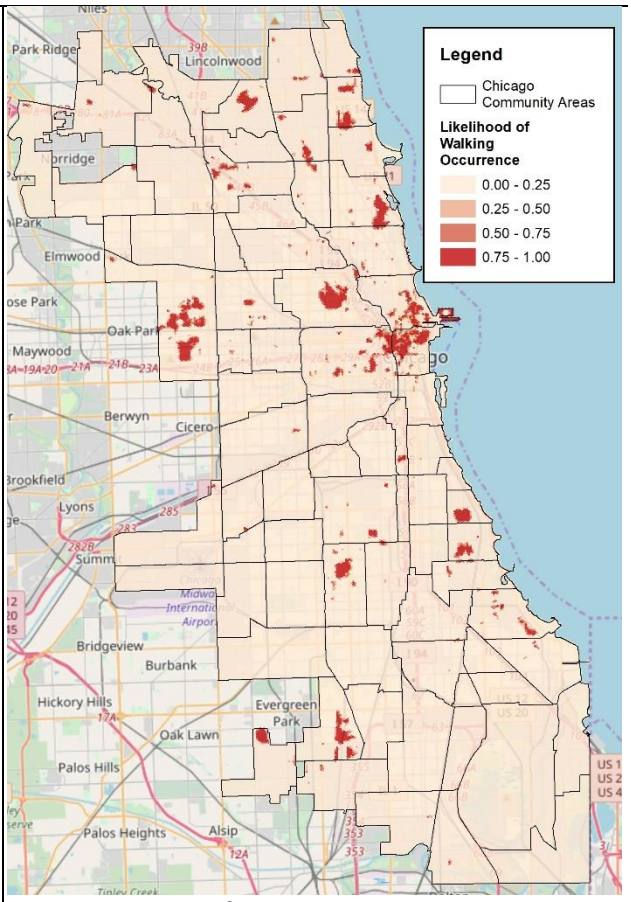
(c) Predicted travel mode occurrence using radial support vector machine

Figure 4.4. Travel mode potential maps across Chicago area, created by trained random forest, extreme gradient boosting, and radial support vector machine classifiers. Predicted walking, in-vehicle status, and biking patterns over Chicago area when (a) random forest, (b) extreme gradient boosting, or (c) radial support vector machine was used.

Figure 4.5. Travel mode likelihood maps in Chicago using RF (100 m hexagon mesh).

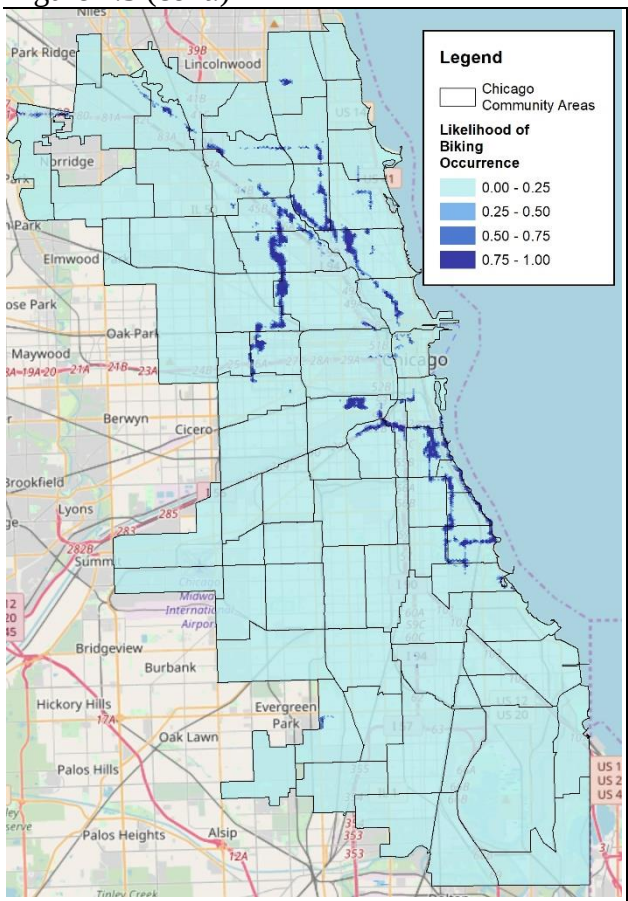


(a) Predicted occurrence of high frequent travel modes

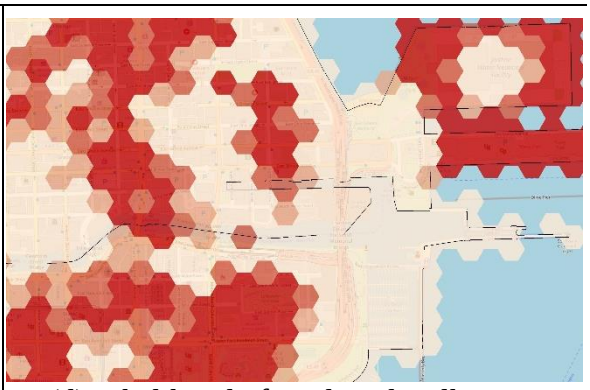


(b) Likelihood of predicted walking occurrence

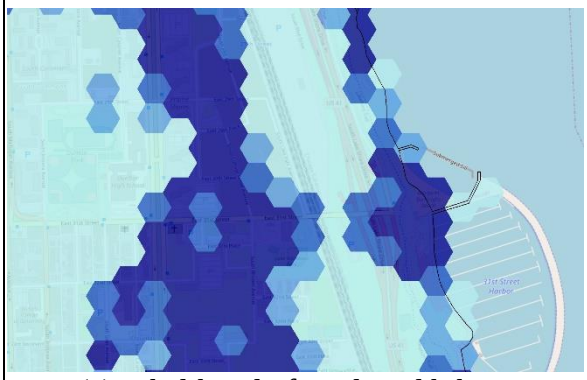
Figure 4.5 (cont.)



(c) Likelihood of predicted biking occurrence



(d) Likelihood of predicted walking in downtown of Chicago



(e) Likelihood of predicted biking in Douglas community

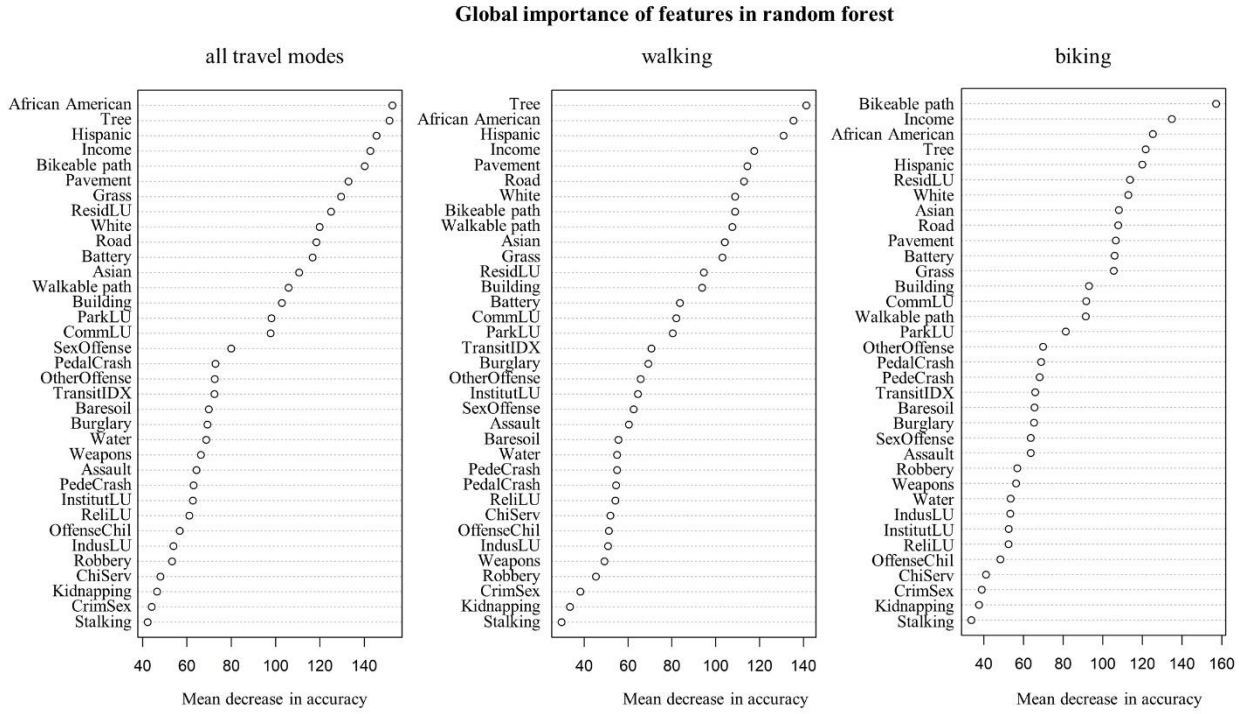


Figure 4.6. The relative importance of 35 features in random forest for all travel modes or active travel modes (walking, biking). Population by ethnic groups (African American, Hispanic, Asian, and White), average neighborhood household income (Income), percentage of natural or built environment area (Road, Building, Tree, Grass, Pavement, Baresoil, Water), total length of walkable or bikeable path per m^2 area (Walkable path, and Bikeable path), average transit accessibility (TransitIDX), number of incidents of different types of crimes per km^2 area (Battery, Assault, OtherOffense, Burglary, Robbery, SexOffense, Weapons, OffenseChil, CrimSex, Stalking, Kidnapping), percentage of land use area (CommLU, IndusLU, InstitutLU, ParkLU, ReliLU, and ResidLU), number of incidents of traffic crashes involving pedestrians or pedal cyclists per km^2 area (PedeCrash, PedalCrash), number of reports regarding abandoned vehicles and graffiti per km^2 area (ChiServ).

The partial dependence (bold line) and centered ICE curves (black) in Figure 4.7 give further insights into some features, including the top-most important features, and their effects on the predictions of three travel modes. Each ICE curve represents the prediction of each observation according to the variation of a selected feature. In African American, the partial dependence showed the positive impact (above zero on the y-axis) on correctly predicting walking as it increases — it reaches to almost 0.1 in y-axis (relatively 10% probability increase of predicting walking) at the highest African American population. On the contrary, African American population was only associated with the predicted biking when it was extremely low under 1,000. On average, walking was also found to have a dependence on the higher tree density showing the partial dependence line begins to increase from zero in y-axis at 30 % of tree in x-axis. On the other hand, ICE curves of biking indicated that low percentage of tree under 30 % affected the predictions of biking. Further, bikeable path only had positive impact on the prediction of biking when it was over 0.003 (m/m^2). As one of crime factors, incidence of battery better predicted walking and in-vehicle status when it was extremely high, while low incidence under 200 count/km^2 was found to be associated with the biking prediction.

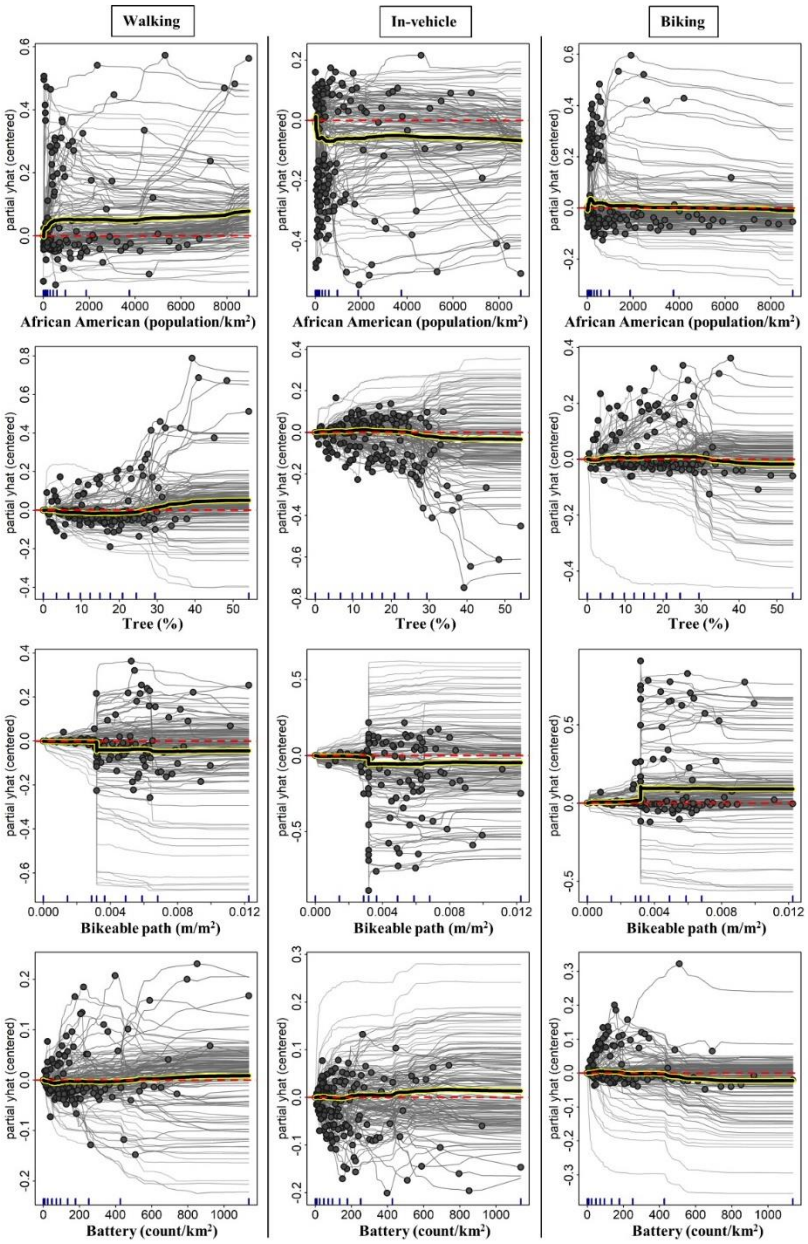


Figure 4.7 Partial dependence and individual conditional expectation of walking (left column), in-vehicle status (middle column), and biking (right column) on neighborhood African American, tree, bikeable path, and crime of battery. Point: an observation of a feature randomly sampled from training data. Gray curve: an individual conditional expectation showing the dependence of travel mode prediction on a feature for each observation. Black-yellow curve: a partial dependence showing average dependence of travel mode prediction on a feature considering all the individual conditional expectations. X-axis: a value range of a feature. Y-axis: a variation of estimated probability of a feature in prediction of a travel mode

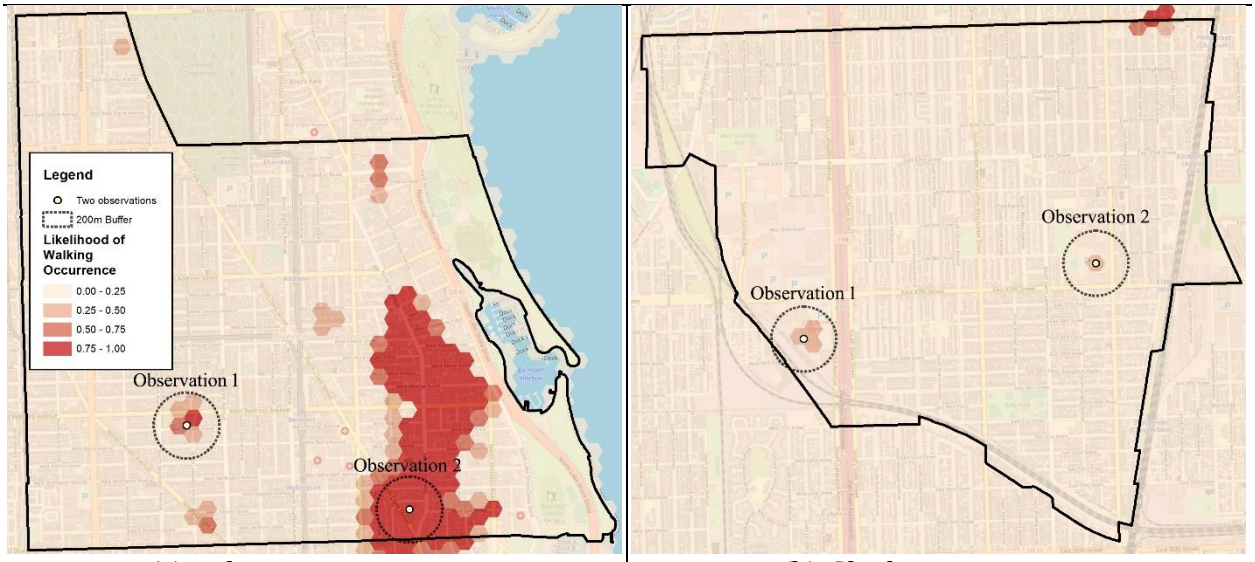
4.4.3 Local impact of environmental contexts on travel modes

One of the unique capabilities of SMAIN is that it facilitates the interpretation of the local impact of environmental contexts on travel modes with the use of explanatory tools. As illustrated in Figure 4.8, the Lakeview community showed highly clustered walking patterns with high likelihood (0.75 – 1.00) in some places such as commercial areas on the east side, whereas a few predicted walking patterns in the Chatham community were present with relatively low likelihood on the west and east sides.

Prediction of walking at the local level could be further explained by exploring the most influential features of each observation (a GPS point represented as a centroid of one of the smallest hexagons with d = 30m) and their contributions to either support or contradict the determination. Figure 4.9 demonstrates the top 10 most important local features of the prediction of walking for two observations in LakeView and another two in Chatham (represented as two points in Figure 4.8 respectively). When compared to the globally important features, the most notable change was that some crime types including assault and offense involving children turned to be important in LakeView and Chatham at a local scale. With regard to the observation 1 in LakeView, moderate to high levels of religious, institutional, and industrial land uses contributed to correct prediction of walking (See Figure B in the Section for value ranges of all features). The observation 2, however, showed that high incidence of traffic crashes (between 38.2 and 50.9 count/km^2) had critical impact on determining walking correctly.

For the other two observations in Chatham, the high percentage of paved areas (more than 31.7 %) better described the correct prediction of walking. The high percentage of bare soil (> 2.045 %) also contributed to the correct prediction for the observation 1. Small percentage of

religious land use less than 0.255 % and high level of parks and open spaces more than 7.13 % also had critical effects on the prediction of walking for the observation 2.



(a) LakeView community (b) Chatham community

Figure 4.8. Exploration of two different communities in Chicago with the likelihood of predicted walking. (a) likelihood of walking occurrence in LakeView community, (b) likelihood of walking occurrence in Chatham community.

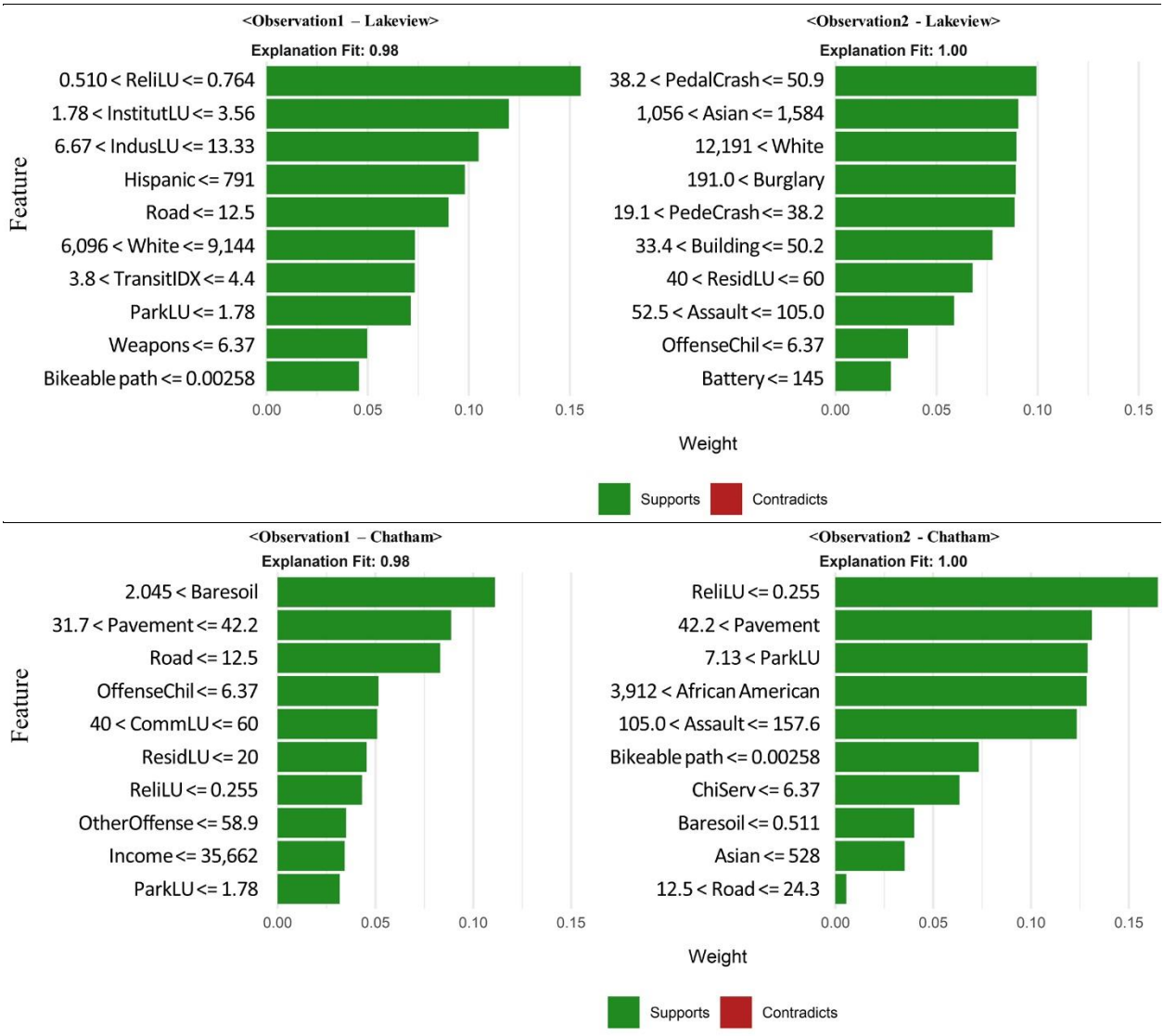


Figure 4.9. Explanatory analysis of the prediction of walking for sampled two observations in Lakeview and Chatham separately associated with top 10 important environmental factors. X-axis: relative importance of features

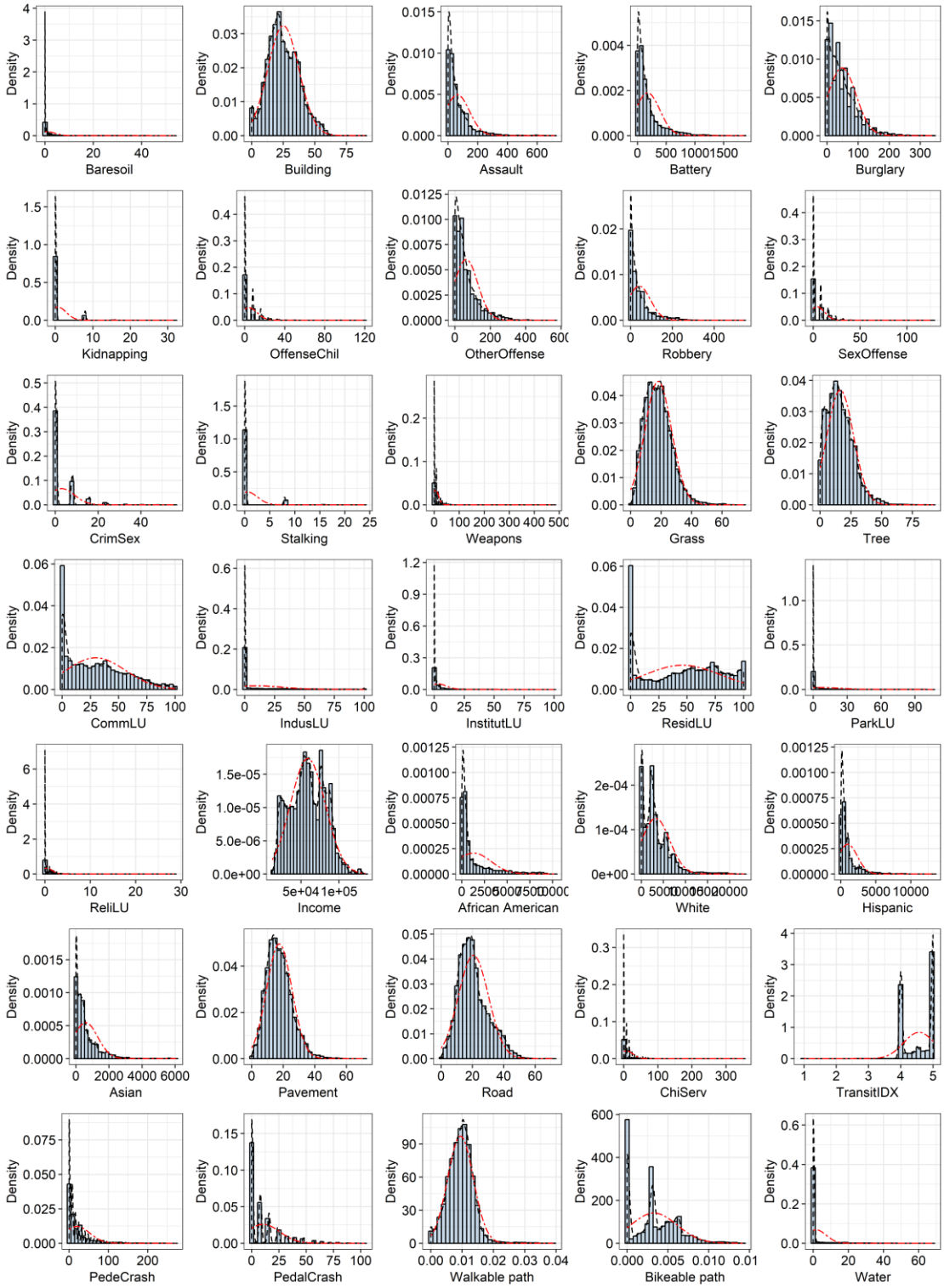


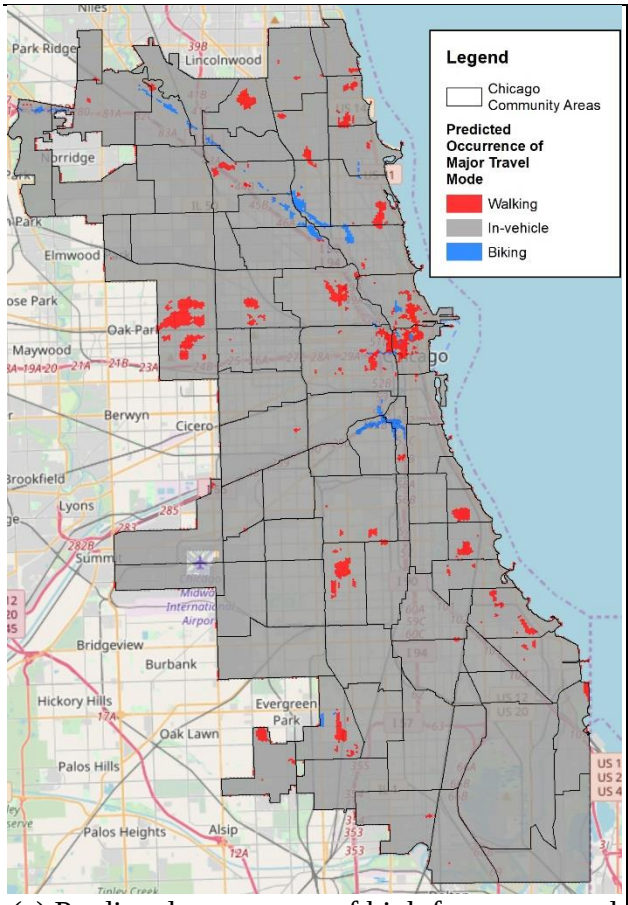
Figure 4.10. Histograms to support the interpretation of effects of 35 environmental factors at local level. Blue-dashed line: density distribution, Red-dashed curve: normal distribution

4.4.4 Temporal dimension of the impact of environmental contexts on travel mode

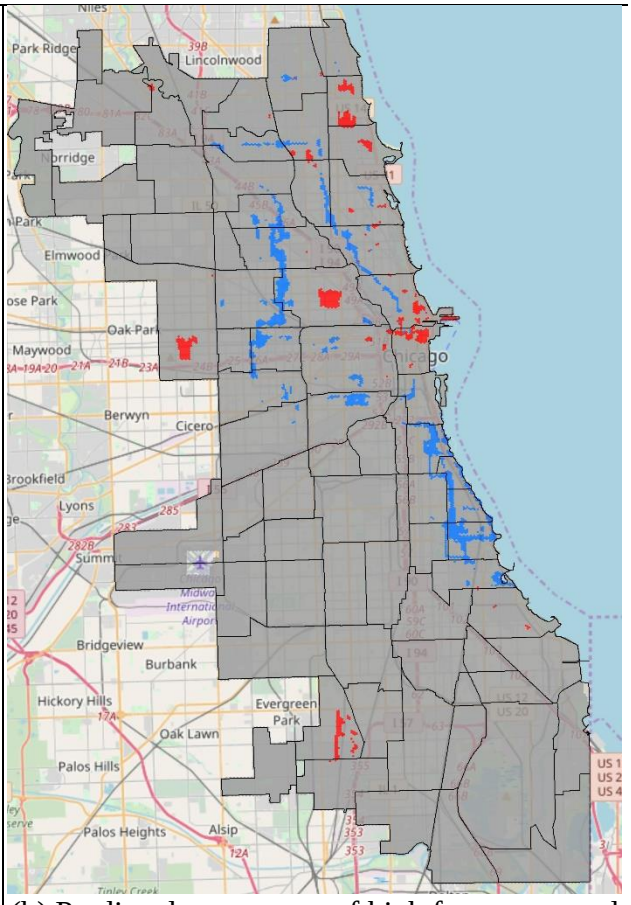
The way in which the temporal dimension influences the impact of environmental contexts on travel modes was visually examined through likelihood mapping by comparing weekday and weekend patterns of ATMs as shown in Figure 4.11 (a), (b). The remarkable difference was that overall, walking and biking patterns during the weekends were found to be concentrated in a few areas, like downtown for walking and lakeside and parks for biking. Moreover, it was revealed that biking during weekends was predicted across much larger areas than weekdays. Interestingly, during weekends, walking was rarely predicted on the south of Chicago when compared to weekdays.

Through the comparison of important features, it was found that neighborhood Asian population had a huge difference of its relative importance between weekdays and weekend days when it contributed to the prediction of walking (Table 4.4). Additional investigation was performed using partial dependence and ICE plots to see how it affected prediction of walking. As shown in Figure 4.11 (c), the contribution of surrounding Asian population in weekend days became much larger than weekdays when it was more than roughly 800 population/km^2. It indicated that when people in Chicago walked, they were exposed to high Asian population during weekends, and it had relatively higher influence on their walking during weekends than weekdays.

Figure 4.11. Maps of predicted occurrence of high frequent travel modes in Chicago during weekdays and weekends and partial dependence and individual conditional expectation of walking on neighborhood Asian population. (a): predicted occurrence of high frequent travel mode during weekdays, (b) predicted occurrence of high frequent travel mode during weekends, (c) Partial dependence and individual conditional expectation of walking on neighborhood Asian population during weekdays (left) and weekend days (right). Point: an observation of a feature randomly sampled from training data. Gray curve: an individual conditional expectation showing the dependence of travel mode prediction on a feature for each observation. Black-yellow curve: a partial dependence showing average dependence of travel mode prediction on a feature considering all the individual conditional expectations. X-axis: a value range of a feature. Y-axis: a variation of estimated probability of a feature in prediction of a travel mode

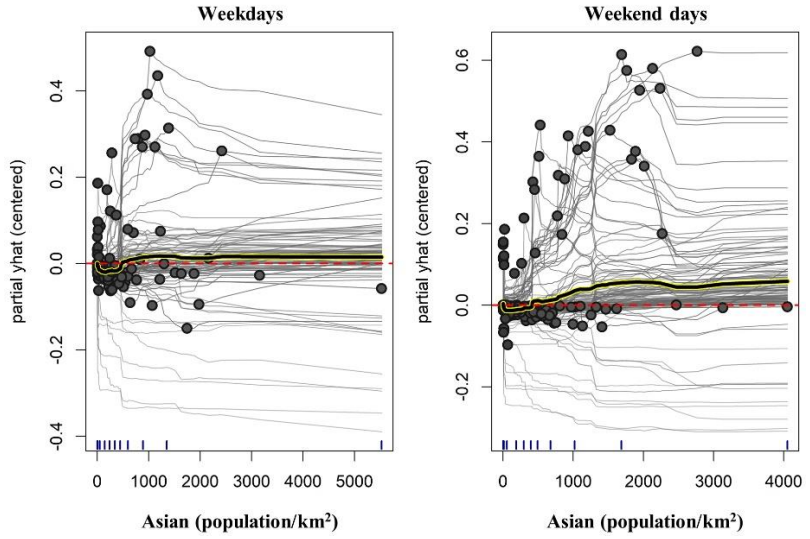


(a) Predicted occurrence of high frequent travel modes during week days



(b) Predicted occurrence of high frequent travel modes during weekends

Figure 4.11 (cont.)



(c)

Table 4.4. Comparison of relatively important features in predicting walking during weekdays and weekend days

Feature	Importance order	
	Weekdays	Weekend days
Tree	1	2
Road	2	5
Walkable path	3	9
Pavement	4	10
White	5	6
Hispanic	6	4
African American	8	1
Asian	11	3

To better understand the contribution of crime factors to the prediction of walking, I further investigated variations of the impact of five crime factors — assault, battery, burglary, other offense, and robbery — on the walking prediction at four different time points — AM peak (6 AM – 8:59 AM), midday (9 AM – 4:59 PM), PM peak (5 PM – 7:59 PM), and night (8 PM –

5:59 AM) — during weekdays and weekends in Chicago. Interestingly, there was a common tendency in ranks of relative importance of the five crime factors indicating that their relative importance mostly increased until the PM peak during a day and immediately dropped at night on weekdays and weekends. The partial dependence and ICE curves provided a way to delve into the associations between those five crime factors and walking prediction at the PM peak during weekdays (Figure 4.13). Battery, other offense, and robbery showed negative associations with the walking prediction to some extent, whereas assault and burglary had positive associations. The negative associations of battery and other offense remained unchanged until they had high incidence levels and yet, those were turned to positive associations when battery and other offense reached to extremely high levels of incidence.

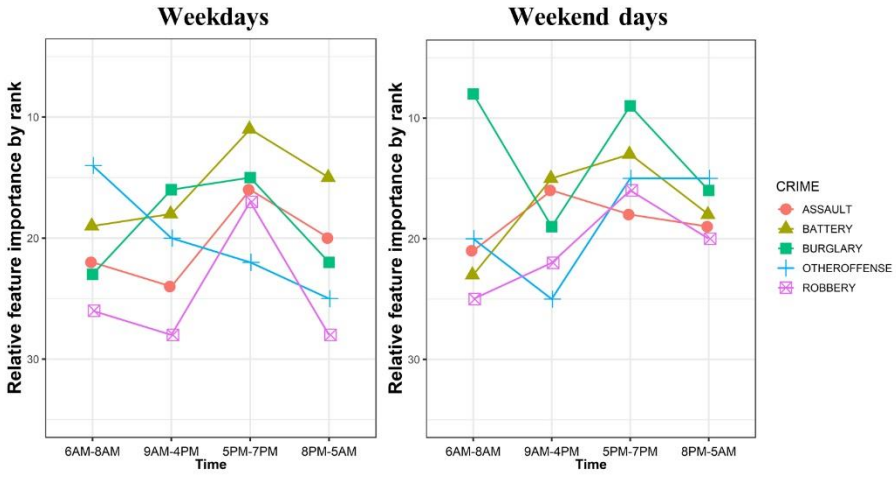


Figure 4.12. Relative importance of assault, battery, burglary, other offense, and robbery by rank at four different time points during weekdays (left) and weekend days (right). Lower values of ranks mean higher feature importance.

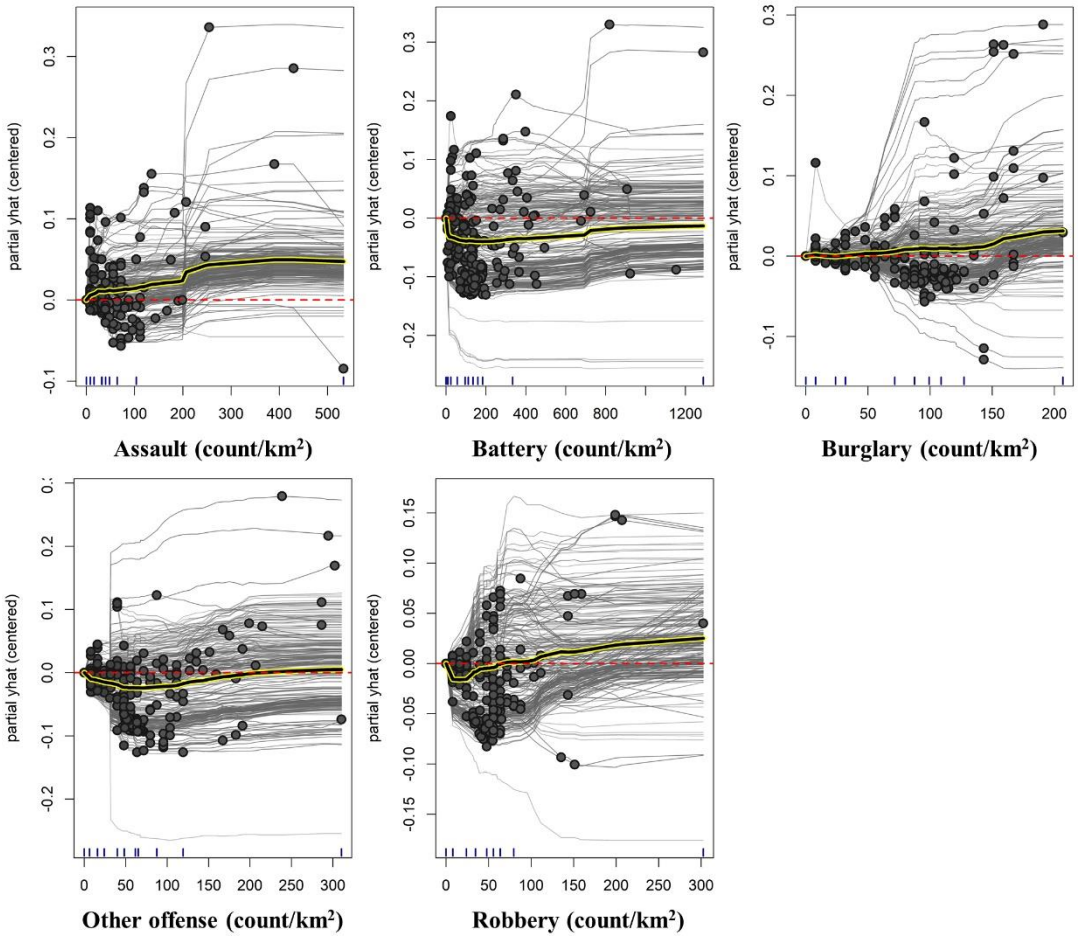


Figure 4.13. Partial dependence and individual conditional expectation of walking on assault, battery, burglary, other offense, and robbery at PM peak during weekdays. Point: an observation of a feature randomly sampled from training data. Gray curve: an individual conditional expectation showing the dependence of travel mode prediction on a feature for each observation. Black-yellow curve: a partial dependence showing average dependence of travel mode prediction on a feature considering all the individual conditional expectations. X-axis: a value range of a feature. Y-axis: a variation of estimated probability of a feature in prediction of a travel mode

4.5 DISCUSSION AND CONCLUSIONS

A novel approach called SMAIN was proposed and adopted to perform likelihood mapping of travel modes in this study. RF was chosen as the best machine learning model with high predictive accuracy of classifying travel modes while considering dynamic exposures to

various environmental contexts along individuals' GPS trajectories. RF especially has inherent strengths in multi-class problems and suited for high-dimensional problems with highly correlated features, like some correlated crime factors in this study (not presented), which might be possible explanations of why RF showed the best predictive accuracy (Gall, Razavi & Van Gool, 2012; Yang et al., 2009). XGB and SVM, however, have a relatively large number of tuning parameters when compared to RF, which made it hard to tune the models appropriately to best predict outcomes. In particular, finding optimal tuning parameters of SVM was limited by the high computational time due to the algorithmic complexity of SVM.

Several likelihood maps for travel modes were created and tailored to portray either the occurrence of major travel modes or the occurrence of each travel mode with the likelihood of occurrence. SMAIN, with the generated likelihood maps and explanatory tools, helped us explore and interpret the complex contextual influences on travel modes globally and locally in a comprehensive view. Further, considering the temporal dimension contributed to understanding how the influences of Asian population and crime factors can vary depending on different periods of time — not only weekdays vs. weekends, but also between different time points within a day. Consequently, with a special focus on the locality and the temporal dimension, this study offered new insights to addressing both UGCoP and spatial non-stationarity through discovering both globally and locally varying contextual influences on travel modes over space and time, which previous studies have not explored. In addition, regarding its applicability, the suggested SMAIN is not limited to a particular human behavior, such as travel modes, but can encompass other kinds of phenomena or events in various domains, including environmental health, to predict them in different regions of interest and examine meaningful patterns that are

hidden behind the large quantity of data while considering human mobility and dynamic interactions between individuals and environments.

In terms of global impact, neighborhood income, racial composition, and tree showed the strongest importance. The neighborhood income and racial composition were the two strongest factors related to walking in previous research (Yu, 2014). With the help of the partial dependence and ICE plots, the percentage of the African-American population around places people passed by was found to be imperative and high level of African American was associated with high likelihood of correct prediction of walking, which was inconsistent with the findings of a previous study that African Americans perceived their neighborhoods as less safe and less pleasant for PA (and thus may lead to lower levels of PA) (Boslaugh et al., 2004). It was, however, revealed that African American had a slightly negative association with biking. The high level of tree density was also found to have strong positive impact on the determination of walking, which corresponded with its consistently positive association with and PA in the Cook County where geographically includes Chicago (Wang, Lee & Kwan, 2018) and specifically walking (Nehme et al., 2016). On the other hand, the low level of tree density was related to the prediction of biking. Considering the African-American population and tree density, even though walking and biking are categorized as ATMs, those two factors showed quite different tendencies towards predictions of the two ATMs. In addition, not to mention the bikeable path, crime of battery did work in a selectively way to the walking and biking predictions. This emphasizes that the categorization based on types of activities provides more ideas of how environmental factors affect people's health behaviors.

With respect to the local impact, public safety factors and land use types became increasingly vital in some places of Lakeview and Chatham communities. Although any

interpretation of the findings at the local level could not be generalizable, explanations of the predictions on individual observations further indicated that the surrounding characteristics of places where high likelihood of walking occurred were quite different in those two communities and even two selected observations in the same community. This, in turn, articulated the existence of spatial non-stationarity in the associations between certain contextual factors and human behaviors, which cannot be explained by global models (Brunsdon et al., 1996).

The examination of temporal variation conveyed further findings on the associations between environmental contexts and walking. This study clearly showed that there was a relatively strong association between Asian population and walking on weekend days rather than weekdays. With the aid of the predicted travel modes during weekdays and weekend days, one of possible explanations is high potential exposures to various ethnic groups during weekends in the recreational and commercial places where adults densely walk, like downtown. Further, the varying impact of crime factors on walking during weekdays and weekends in Chicago was one of our pivotal findings that can contribute to the understanding of temporal variations in contextual effects and to addressing the temporal uncertainties associated with the UGCoP. Breaking down crime into different types also helped disclose how each type of crime has influence on the determination of walking, which past studies on perceived or objective crime measures and outdoor PA, have not considered (Saelens et al., 2003; Gomez et al., 2004; Ruijsbroek et al., 2015; Chaudhury et al., 2016). Assault, battery, burglary, other offense, and robbery were found to have relatively high importance in the prediction of walking between 5 and 7 PM during a day on both weekdays and weekends. Among them, battery, other offense, and robbery particularly produced expected findings on weekdays regardless of the fact that the findings of those three crime factors did not have even linear relationships with the predictions of

walking. There might be some reasons of the implausible findings showing consistently positive associations with walking when those five crime factors had extremely high incidence levels against our willingness to enjoy walking in safe places. Thus, more in-depth research need to be followed in order to understand the veiled complexities.

SMAIN showed diverse capabilities of understanding complex geographic and social problems by considering many environmental factors at once and partial dependence and centered ICE plots and LIME as explanatory tools. The likelihood maps allowed visual exploration and further examination of the patterns of interest in different neighborhoods of an entire study area with the help of machine learning techniques. However, this study has several limitations. First, the demographic and socioeconomic characteristics of the sample, such as race, and household income, are not very similar to the actual adult population characteristics of Chicago. The GPS trajectories of the participants in the sample thus may not represent the true or typical travel mode patterns of adults in Chicago. Its implications on the interpretation of the expected ATMs in different neighborhoods with environmental factors regarding geographic and social problems can also be affected. Second, this study does not explain much about the combined effects of contextual factors. Further analysis of the joint impact of two or more influential factors on travel modes can provide more insights into explanations of complex interactions between travel modes and their influential factors. In addition, presumably due to the age of the samples (mostly middle-aged adults), the resulted likelihood maps also had many in-vehicle patterns predicted across Chicago when compared to walking or biking. Thus, young adults who engage in more diverse activities in their daily lives will work as suitable samples for the next step. Last, walking and biking are the only ATMs that were considered in this study, and running was excluded due to the lack of substantial observations. Running is one of the most

popular types of outdoor PA in addition to biking in the U.S. (Outdoor Foundation, 2016). The purposes of walking and biking, which can contribute to an in-depth understanding of the associations between specific environments and recreational and utilitarian travels, were also not examined in this study.

For further work, appropriate sampling to choose a set of participants representative of the large population needs to be performed. To provide more solid interpretations, more restricted demographic and socio-economic groups are necessary to be targeted for more concrete and clearer interpretations of the results. Since this study focused more on demonstrating the various capabilities of likelihood mapping, addressing specific substantive questions for particular segments of the population will need to be conducted in future research. To obtain more generalizable knowledge, participants in different societal and cultural contexts (e.g., other cities or countries) will also need to be considered. In addition, the likelihood mapping technique should be further developed to delineate clearer geographic boundaries for the prediction of travel modes. As represented by the regularly distributed points, this study assumes that the tracks of people are all over the place in Chicago, including water body and abandoned areas. Thus, some constraints will be helpful for accurately assessing possible areas where people can visit and leaving out areas where any of the predicted results cannot be expected. For example, the examination of different delineation methods beyond the buffers are also needed to precisely delineate the neighborhood areas (e.g. Kwan et al., 2018). Some of the built environment factors should be also further divided to produce more meaningful findings since the occurrence of walking found around roads, buildings, and walkable paths is highly predictable, although these factors serve as essential predictors to achieve a high predictive accuracy of travel modes. Accordingly, the subdivided built environment types coupled with

more contextual information (e.g. office buildings, high quality of parks and open spaces) will increase analytical power to resolve more problems in mobility research.

4.6 REFERENCES

Almanza, E., Jerrett, M., Dunton, G., Seto, E. & Pentz, M. A. (2012). A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. *Health & place* **18** (1), 46–54.

Birch, C. P., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological modelling* **206** (3–4), 347–359.

Bohra, A. & Andrianasolo, H. (2001). Application of GIS in modeling of dengue risk based on sociocultural data: Case of Jalore, Rajasthan, India. *Dengue Bulletin* **25**, 92–102.

Boruff, B. J., Nathan, A. & Nijɲnstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood'in place-based health research. *International journal of health geographics* **11** (1), 22.

Boslaugh, S. E., Luke, D. A., Brownson, R. C., Naleid, K. S. & Kreuter, M. W. (2004). Perceptions of neighborhood environment for physical activity: Is it "who you are" or "where you live?". *Journal of Urban Health* **81** (4), 671–681.

Browning, M., & Lee, K. (2017). Within what distance does "Greenness" best predict physical health? A systematic review of articles with GIS buffer analyses across the lifespan. *International journal of environmental research and public health* **14** (7), 675.

Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* **28** (4).

Bui, D. T., Ho, T.-C., Pradhan, B., Pham, B.-T., Nhu, V.-H. & Revhaug, I. (2016a). GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks.

Environmental Earth Sciences **75** (14).

Bui, D. T., Le, K.-T., Nguyen, V., Le, H. & Revhaug, I. (2016b). Tropical Forest Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, Using GIS-Based Kernel Logistic Regression. *Remote Sensing* **8** (4), 347.

Bureau of Justice Statistics. (2018). Violent crime. Retrieved from https://www.bjs.gov/index.cfm?ty=tp&tid=31

Cerin, E., Mitős, J., Cain, K. L., Conway, T. L., Adams, M. A., Schofield, G., Sarmiento, O. L., Reis, R. S., Schipperijn, J., Davey, R., Salvo, D., Orzanco-Garralda, R., Macfarlane, D. J., De Bourdeaudhuij, I., Owen, N., Sallis, J. F. & Van Dyck, D. (2017). Do associations between objectively-assessed physical activity and neighbourhood environment attributes vary by time of the day and day of the week? IPEN adult study. *International Journal of Behavioral Nutrition and Physical Activity* **14** (1), 34.

Chang, A. Y., Parrales, M. E., Jimenez, J., Sobieszczyk, M. E., Hammer, S. M., Copenhaver, D. J. & Kulkarni, R. P. (2009). Combining Google Earth and GIS mapping technologies in a dengue surveillance system for developing countries. *International Journal of Health Geographics* **8** (1), 49.

Chaudhury, H., Campo, M., Michael, Y. & Mahmood, A. (2016). Neighbourhood environment and physical activity in older adults. *Social Science & Medicine* **149**, 104–113.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research* **16** (1), 321–357.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of*

the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Chicago Data Portal. (2013). Selected socioeconomic indicators in Chicago, 2007 – 2011. Retrieved from https://data.cityofchicago.org/Health-Human-Services/below-poverty-level-by-community/b7zw-zvm2

Chicago Metropolitan Agency for Planning (CMAP) Data Hub (2017). Transit availability index. Retrieved from https://datahub.cmap.illinois.gov/dataset/access-to-transit-index

Chicago Police Department. (2007). 2007 annual report. Retrieved from https://home.chicagopolice.org/wp-content/uploads/2014/12/2007-Annual-Report.pdf

Cooper, A. R., Page, A. S., Wheeler, B. W., Griew, P., Davis, L., Hillsdon, M. & Jago, R. (2010). Mapping the walk to school using accelerometry combined with a global positioning system. *American journal of preventive medicine* **38** (2), 178–183.

Delmelle, E. M., & Delmelle, E. C. (2012). Exploring spatio-temporal commuting patterns in a university environment. *Transport Policy* **21**, 1–9.

Dodge, S. (2016). From Observation to Prediction: The Trajectory of Movement Research in GIScience.. In H. Onsrud & W. Kuhn (ed.),*Advancing Geographic Information Science: The Past and Next Twenty Years* (pp. 123). GSDI Association Press.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29** (5), 1189–1232.

Gall, J., Razavi, N. & Van Gool, L. (2012). An Introduction to Random Forests for Multi-class Object Detection. In F. Dellaert, J.-M. Frahm, M. Pollefeys, L. Leal-Taixй & B. Rosenhahn (ed.), *Outdoor and Large-Scale Real-World Scene Analysis* (pp. 243–263). Springer Berlin Heidelberg.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* **24** (1), 44–65.

Gomez, J. E., Johnson, B. A., Selva, M. & Sallis, J. F. (2004). Violent crime and outdoor physical activity among inner-city youth. *Preventive medicine* **39** (5), 876–881.

Hoehner, C. M., Ramirez, L. K. B., Elliott, M. B., Handy, S. L. & Brownson, R. C. (2005). Perceived and objective environmental measures and physical activity among urban adults. *American journal of preventive medicine* **28** (2), 105–116.

Jansen, M., Ettema, D., Pierik, F. & Dijst, M. (2016). Sports Facilities, Shopping centers or homes: What locations are important for adults' physical activity? A cross-sectional study. *International journal of environmental research and public health* **13** (3), 287.

Keddem, S., Barg, F. K., Glanz, K., Jackson, T., Green, S. & George, M. (2015). Mapping the urban asthma experience: Using qualitative GIS to understand contextual factors affecting asthma control. *Social Science & Medicine* **140**, 9–17.

King, A. C., Stokols, D., Talen, E., Brassington, G. S. & Killingsworth, R. (2002). Theoretical approaches to the promotion of physical activity: forging a transdisciplinary paradigm. *American journal of preventive medicine* **23** (2), 15–25.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697). ACM.

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers* **102** (5), 958–968.

Kwan, M.-P. (2013). Beyond space (as we knew it): Toward temporally integrated geographies

of segregation, health, and accessibility. *Annals of the Association of American Geographers* **103** (5), 1078–1086.

Kwan, M.-P. (2018a). The limits of the neighborhood effect: Contextual uncertainties in geographic, environmental health, and social science research. *Annals of the American Association of Geographers*, **108** (6), 1482–1490.

Kwan, M.-Po. (2018b). The neighborhood effect averaging problem (NEAP): An elusive confounder of the neighborhood effect. *International Journal of Environmental Research and Public Health*, **15,** 1841.

Kwan, M.-P., Wang, J., Tyburski, M., Epstein, D.H., Kowalczyk, W.J., & Preston, K.L. (2018). Uncertainties in the geographic context of health behaviors: A study of substance users' exposure to psychosocial stress using GPS data. *International Journal of Geographical Information Science*, forthcoming.

Lachowycz, K., Jones, A. P., Page, A. S., Wheeler, B. W. & Cooper, A. R. (2012). What can global positioning systems tell us about the contribution of different types of urban greenspace to children's physical activity?. *Health & place* **18** (3), 586–594.

Lee, K and Kwan, M.-P. (2018). Automatic physical activity and in-vehicle status classification based on GPS and accelerometer data: A hierarchical classification approach using machine learning techniques. *Transactions in GIS*, forthcoming.

Lee, S., Hong, S.-M. & Jung, H.-S. (2017). GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. *Geocarto International* **33** (8), 847–861.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of*

Applied Statistics, **9** (3), 1350–1371.

McGinn, A. P., Evenson, K. R., Herring, A. H., Huston, S. L. & Rodriguez, D. A. (2007). Exploring associations between physical activity and perceived and objective measures of the built environment. *Journal of Urban Health* **84** (2), 162–184.

McNeill, L. H., Kreuter, M. W. & Subramanian, S. (2006). Social environment and physical activity: a review of concepts and evidence. *Social science & medicine* **63** (4), 1011–1022.

Miranda, M. L., Dolinoy, D. & Alicia Overstreet, M. (2002). Mapping for Prevention: GIS Models for Directing Childhood Lead Poisoning Prevention Programs. *Environmental Health Perspectives* **110**, 947–953.

Nagel, C. L., Carlson, N. E., Bosworth, M. & Michael, Y. L. (2008). The relation between neighborhood built environment and walking activity among older adults. *American journal of epidemiology* **168** (4), 461–468.

Naghibi, S. A., Pourghasemi, H. R. & Dixon, B. (2015). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment* **188** (1).

National Institute of Justice. (2018). Violent crimes. Retrieved from https://www.nij.gov/topics/crime/violent/Pages/welcome.aspx

Nehme, E. K., Oluyomi, A. O., Calise, T. V., & Kohl, H. W. (2016). Environmental Correlates of Recreational Walking in the Neighborhood. *American Journal of Health Promotion*, **30** (3), 139–148.

Outdoor Foundation (2016). Outdoor recreation participation topline report 2016. Retrieved from

http://www.outdoorfoundation.org/pdf/ResearchParticipation2016Topline.pdf.

Park, Y. M. & Kwan, M.-P. (2017). Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health & place* **43**, 85–94.

Perchoux, C., Chaix, B., Cummins, S. & Kestens, Y. (2013). Conceptualization and measurement of environmental exposure in epidemiology: accounting for activity space related to daily mobility. *Health & place* **21**, 86–93.

Physical Activities Guidelines Advisory Committee. (2008). Physical activity guidelines advisory committee report. Washington (DC): US Department of Health and Human Services.

van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* **14** (1), 137.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences* **51**, 350–365.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Rigolon, A., Browning, M. H. E. M., Lee, K. & Shin, S. (2018). Access to Urban Green Space in Cities of the Global South: A Systematic Literature Review. *Urban Science* **2** (3).

Rodriguez, D. A., Cho, G.-H., Evenson, K. R., Conway, T. L., Cohen, D., Ghosh-Dastidar, B., Pickrel, J. L., Veblen-Mortenson, S. & Lytle, L. A. (2012). Out and about: association of

the built environment with physical activity behaviors of adolescent females. *Health & place* **18** (1), 55–62.

Roux, A. V. D. & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences* **1186** (1), 125–145.

Ruijsbroek, A., Droomers, M., Groenewegen, P. P., Hardyns, W. & Stronks, K. (2015). Social safety, self-rated general health and physical activity: changes in area crime, area safety feelings and the role of social cohesion. *Health & place* **31**, 39–45.

Ruiz, M. O., Tedesco, C., McTighe, T. J., Austin, C. & Kitron, U. (2004). Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. *International Journal of Health Geographics* **3** (1), 8.

Ruiz, M. O., Chaves, L. F., Hamer, G. L., Sun, T., Brown, W. M., Walker, E. D., Haramis, L., Goldberg, T. L. & Kitron, U. D. (2010). Local impact of temperature and precipitation on West Nile virus infection in Culex species mosquitoes in northeast Illinois, USA. *Parasites & vectors* **3** (1), 19.

Saelens, B. E., Sallis, J. F., Black, J. B. & Chen, D. (2003). Neighborhood-Based Differences in Physical Activity: An Environment Scale Evaluation. *American Journal of Public Health* **93** (9), 1552–1558.

Sallis, J., Bauman, A. & Pratt, M. (1998). Environmental and policy interventions to promote physical activitya. *American journal of preventive medicine* **15** (4), 379–397.

Sallis, J. F., Floyd, M. F., Rodriguez, D. A. & Saelens, B. E. (2012). Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* **125** (5), 729–737.

Sallis, J. F., Owen, N. & Fisher, E. (2015). Ecological models of health behavior. *Health behavior: Theory, research, and practice* **5**, 43–64.

Song, C., Kwan, M.-P., Song, W. & Zhu, J. (2017). A Comparison between spatial econometric models and random forest for modeling fire occurrence. *Sustainability* **9** (5), 819.

Spence, J. C. & Lee, R. E. (2003). Toward a comprehensive model of physical activity. *Psychology of sport and exercise* **4** (1), 7–24.

Stewart, C. A., Cockerill, T. M., Foster, I., Hancock, D., Merchant, N., Skidmore, E., Stanzione, D., Taylor, J., Tuecke, S., Turner, G., Vaughn, M. & Gaffney, N. I. (2015). Jetstream: A Self-provisioned, Scalable Science and Engineering Cloud Environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* (pp. 29:1–29:8). ACM.

Tehrany, M. S., Pradhan, B. & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of Hydrology* **512**, 332–343.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R. & Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering* **16** (5), 62–74.

Troped, P. J., Wilson, J. S., Matthews, C. E., Cromley, E. K. & Melly, S. J. (2010). The built environment and location-based physical activity. *American journal of preventive medicine* **38** (4), 429–438.

United States Census Bureau (2010). 2010 census summary of Chicago city, Illinois. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

Wang, J., & Kwan, M. P. (2018a). An analytical framework for integrating the spatiotemporal

dynamics of environmental context and individual mobility in exposure assessment: A study on the relationship between food environment exposures and body weight. *International journal of environmental research and public health* **15** (9), 2022.

Wang, J., & Kwan, M.-P. (2018b). Hexagon-based adaptive crystal-growth Voronoi diagrams based on weighted planes for service area delimitation. *ISPRS International Journal of Geo-Information* **7** (7), 257.

Wang, J., Lee, K. & Kwan, M.-P. (2018). Environmental influences on leisure-time physical inactivity in the US: An exploration of spatial non-stationarity. *ISPRS International Journal of Geo-Information* **7** (4), 143.

Yang, F., Wang, H. Z., Mi, H., Lin, C. D., & Cai, W. W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics* **10** (1), S22.

Yu, C. Y. (2014). Environmental supports for walking/biking and traffic safety: income and ethnicity disparities. *Preventive medicine* **67**, 12–16.

Zabihi, M., Pourghasemi, H. R., Pourtaghi, Z. S. & Behzadfar, M. (2016). GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environmental Earth Sciences* **75** (8).

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

5.1 SUMMARY

In this dissertation, I proposed how to investigate the associations between travel modes and various environmental factors from the perspective of data-driven approaches. In the era of big data, I suggested methodological directions for various fields of study to adequately deal with a large quantity of sensor data collected from many participants, derive informative features for classifying health behaviors from the sensor data, and conduct exploratory analyses and produce meaningful knowledge using GIS data. One of the key techniques tying the series of methodological improvements together, introduced in this dissertation, was machine learning as a branch of artificial intelligence. In Chapter 2, I developed a novel approach to automatic classification of travel modes using published GPS and accelerometer data. With the predicted travel modes, I showed how buffer analysis, which had been widely used in many physical activity (PA) and transportation studies to estimate exposures to ambient environments, could affect the associations between travel modes and various environmental factors, with a special focus on its size in Chapter 3. In Chapter 4, to take advantage of a large amount of individuals' daily movement data, complex associations between travel modes and 35 environmental factors were examined using machine learning models rather than conventional statistical models, which are not very suitable to discover hidden meaningful patterns from the large quantity of daily movement data with many predictors. Supercomputing resources were an integral part for efficient buffer analysis to assess contextual influences considering all GPS points on individual paths. An innovative framework called Spatio-temporal Mapping And INterpretation (SMAIN) was proposed to formalize the accompanying processes, from data processing to interpretation of findings.

5.2 FINDINGS AND CONTRIBUTIONS

This dissertation provided and addressed empirical evidence and multiple methodological aspects for health, transportation, and urban planning research. First, it was found that the racial composition and economic status that people are exposed to in proximate environments were critical factors to explain what kinds of travel modes people are engaged in. Those two neighborhood characteristics were the most critical features to determine people's travel modes, suggesting that demographic information and socio-economic status at both individual and neighborhood levels need to be considered together in ambient environmental assessments. In addition, as a proxy of greenness, the tree density was one of the environmental determinants that showed consistent outcomes in terms of the prediction of walking, whereas parks and open spaces showed mixed findings or did not have a huge bearing on walking. For safety-related factors, battery, other offenses, and robbery showed consistently negative associations with walking to some extent at PM peak on weekdays with relatively high importance, while assault and burglary had unexpected outcomes. The unexpected findings in particular crime types might reflect the potential gap between the objective measures used in this study and individual perceptions of how they feel safety in practice (Centers for Disease Control and Prevention, 1999; Brownson et al., 2000).

This dissertation research attempted to address several methodological issues that may increase accurate travel mode detection, contextual influence assessments, and investigation of associations between human behaviors and environmental contexts. First, the developed automatic classification contributes to mobility research in various research domains, including health and transportation. The objective and accurate travel mode detection using heterogeneous

sensor data helps prevent any missing or mistaken records in surveys, which might vary outcomes of research. This study serves as guidance for automatic classification of travel modes for researchers to replicate the same approach with the published sensor data and further modify and improve it depending on the scope of research.

Second, considering the daily movement of individuals using GPS data, this research addressed uncertain geographic context problem (UGCoP) in GIS research to mitigate spatial and temporal uncertainties that may considerably affect the assessments of contextual influence. With the suggested SMAIN framework, the uncertainties in terms of the interactions between individuals and environments in specific areas and time points were alleviated. It especially supplemented conventional statistical models, demonstrating clear interactions between each travel mode and crime factor at different time points on weekdays and weekend days, using partial dependence and individual conditional expectation (ICE) plots. Further, through sensitivity analyses, some factors, including crime, appeared to be sensitive to the size of buffers. Kwan (2012) and Shi (2010) especially emphasized the vital role of sensitivity analyses in understanding how delineations of contextual units exert influence on research findings. Larger sizes, such as 150 m and 200 m, were found to have more significance levels in given environmental factors when logistic regression models were used. One of the most important findings was that the buffer-size effect was alleviated when a model included critical variables, like neighborhood income, that could greatly improve its fit.

Third, the spatial non-stationarity was also addressed by creating global and local models and observing how the contributions of environmental factors were presented between the two different scales of models. Local Interpretable Model-agnostic Explanations (LIME), as a useful explanatory tool, provided opportunities to magnify local characteristics and explore elusive

patterns when machine learning algorithms were applied to build models. For instance, different types of land use and crimes not included in the group of most influential features at the global model turned out to have a great influence on predicting walking in the two selected communities in Chicago.

5.3 IMPLICATIONS

Through the statistical and machine learning approaches, the impact of social and physical contexts in immediate surroundings where people's active travels are performed was understood well, focusing on an urban area. The knowledge gained from the local perspective, especially, gave more insights into what contexts in different communities could affect people's daily movements. Consequently, the findings from the local characteristics could give ideas of how local governments should establish strategies targeting specific contexts to promote active travels. For example, the high percentage of paved surfaces that better predicted walking with relatively high importance in the Chatham community in Chicago means that the walking environment is not pleasant and needs to be improved. In this way, findings at the finer level can provide more descriptions to portray locally important contexts in a specific place for healthy community design, in addition to more general knowledge gained through global models.

Tree density as one of the most influential factors showing consistent outcomes is the foremost built environment that may need to be considered for encouraging adults in communities to walk. Brisk walking, especially, brings us health benefits and makes communities active and healthy. Hence, building walking-friendly environments is a foundational and essential part in community design. To promote walking, a wide distribution of

walking paths landscaped with trees should be taken into account, which can also improve the aesthetics of the community environment.

Moreover, this dissertation research suggests targeting a specific active travel for policies and interventions in urban planning and design. The results indicated that several environmental factors, including tree density, did not exert the same effects on walking and biking. The high percentage of tree density more than 30 % was found to have a great influence on walking, whereas tree density less than 30 % was likely to correctly predict biking. Moderate to high levels of bikeable paths (more than 0.003 m/m^2) made a great contribution to biking. Thus, when urban planners come up with a blueprint to redesign communities, they may need to consider such selective effects of contexts for the promotion of active travels.

5.4 FUTURE WORK

For future research, the error propagation of the predicted travel modes should be investigated. One major concern regarding the use of predicted travel modes is that inevitable errors of the prediction may affect research findings of the associations between travel modes and environmental factors. To minimize such wrong prediction, I strove to improve the travel mode classification algorithm, considering slow traffic due to severe traffic congestions in urban areas. The enhanced approach, however, failed to filter out the bad quality of GPS points, which could generate wrong predictions, especially downtown with tall buildings since there was no indicator regarding horizontal accuracy recorded in the GPS data collected through Chicago Regional Household Travel Inventory (CRHTI). Therefore, more systematic analyses need to be conducted to explore whether study outcomes greatly vary according to different scenarios in terms of the availability of GPS attributes.

More in-depth investigation of buffers could be performed to address the methodological issue of the definition of accurate contextual units to find true geographic knowledge. The role of buffers in the assessment of dynamic, instant, and immediate surroundings was significant; however, there are some variants of buffer developed in previous studies (Boruff et al., 2012); thus, taking those approaches into consideration will be also valuable to precisely define proximate areas where environmental exposures have a great influence on people's behavior.

The mixed associations between active travel modes (ATMs) and some safety-related factors, like traffic crashes, still remain a challenging issue. Following research will need to focus on their perceived aspects regarding people to delve into how the perceived levels of safety-related factors affect people's ATMs. In addition, both the efficacy of perceptions and objective measures of public safety will be evaluated to give insights into how those two measurements are linked to people's ATMs.

Since this research is full of quantitative analyses, there is a paucity of qualitative viewpoints on the associations between travel modes and environmental factors. Qualitative information can provide rationales for the mechanism of interactions of people with environments shaped in a certain community or a larger region. For instance, the mixed findings of crime and traffic crashes can be further examined through in-depth interviews to know what people in certain communities usually feel regarding safety when they walk, and what factors of crime and traffic crashes they think matter with their active travels. Accessibility of parks and open spaces can be also assessed by the qualitative approach (e.g., assessment of park quality involving amenities and barriers) to account for the reasons why parks and open spaces act as a barrier to perform walking and biking for some groups of people — middle-aged white adults in this study.

5.5 REFERENCES

Boruff, B. J., Nathan, A., & Nijënstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood'in place-based health research. *International journal of health geographics*, *11*(1), 22.

Brownson, R. C., Housemann, R. A., Brown, D. R., Jackson-Thompson, J., King, A. C., Malone, B. R., & Sallis, J. F. (2000). Promoting physical activity in rural communities: walking trail access, use, and effects. *American journal of preventive medicine*, *18*(3), 235-241.

Centers for Disease Control and Prevention (CDC. (1999). Neighborhood safety and the prevalence of physical inactivity--selected states, 1996. *MMWR. Morbidity and mortality weekly report*, *48*(7), 143.

Kwan, M. P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, *102*(5), 958-968.

Shi, W. (2010). Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses. Boca Raton: CRC Press, https://doi.org/10.1201/9781420059281